LAUREA AMMATTIKORKEAKOULU Uuden edellä

Tietovarastointi ja liiketoimintatiedon hallinta – Case TSF

Saastamoinen, Joni

2012 Leppävaara

Laurea-ammattikorkeakoulu
Laurea Leppävaara

Tietovarastointi ja liiketoimintatiedon hallinta – Case TSF

Saastamoinen Joni
Tietojenkäsittelyn koulutusohjelma
Opinnäytetyö
marraskuu, 2012

Laurea-ammattikorkeakoulu **Tiivistelmä**
Laurea Leppävaara
Tietojenkäsittelyn koulutusohjelma

Saastamoinen Joni

Tietovarastointi ja liiketoimintatiedon hallinta – Case TSF

Vuosi 2012 Sivumäärä 51

Tässä opinnäytetyössä tutustutaan yritysmaailman tietovarasto- (Data Warehouse, DW) ja liiketoimintatiedon hallinnan (Business Intelligence, BI) tarpeisiin sekä mitä niillä oikein toteutettuna voidaan saavuttaa. Opinnäytetyön tekijä suorittaa syventävän työharjoittelun sekä tekee opinnäytetyön TeliaSonera Finlandissa, Group IT, Business Intelligence & Data Warehouse (BI&DW) -yksikössä.

Tietovarastointi ja liiketoimintatiedon hallinta ovat nopeimmin kehittyviä tietotekniikan alueita. Monet yritykset ja julkishallinnon toimijat ovat huomanneet, että operatiiviset järjestelmät eivät pelkästään itsessään riitä palvelemaan tarpeeksi hyvin vaativia tietojen raportointi- ja analysointitarpeita. Informaation jatkuva määrällinen kasvu ja tallennus tulee hallita kokonaisuudessaan järjestelmällisesti. Näin saadaan jalostettua tärkeää informaatiota – uutta, vanhaa sekä ennakoivaa, toimijoiden päätöksenteon tueksi.

Yksi tavoitteista oli selvittää, löytyykö PWCIMATICA (PowerCenter Informatica) - repositoriosta samanlaista metadataa kuin yrityksen ETL (extract, transform and load) -työkalu Fileloader tuottaa. Repositoriosta löytyi hyvin Fileloaderin tuottaman metadatan kaltaista metadataa, joskaan kaikkia vastaavuuksia ei repositoriosta löytynyt. Kaikesta metadatasta ei voinutkaan löytyä täysin vastaavaa metadataa, koska osa metadatasta populoidaan käsin.

Lähdejärjestelmien tiedostojen latausstatistiikat löytyivät suurimmaksi osaksi sekä lähdetiedostojen rakennekuvaukset löytyivät. Työtä tehdessä löydettiin myös metadataa, jota yritys voi halutessaan hyödyntää seuratessaan latausprosesseja ja käyttää vikatilanteiden selvittämisessä sekä itse korjauksessa. Myös kohde metadataa löytyi hyvin.

DW&BI ovat käsitteiltään laajoja ja pitävät sisällään monia tehtäviä. DW -ympäristön rakentaminen ja ylläpito isoimmissa yrityksissä onkin kokonaisen osaston vastuulla ja vaatii jatkuvaa tilannetietoisuutta koko osastolta. Tärkeää on myös nimenomaan ylläpidon yhteydenpito BI-käyttäjiin jo suunnitteluvaiheessa, jotta saadaan kaikille asianomaisille tietoisuutta aiheesta. Ympäristön antamien mahdollisuuksien ollessa työntekijöiden tiedossa hyötyvät yritykset DW&BI –toiminnasta merkittävästi, ja se näkyy varmasti yrityksen toiminnassa.

Asiasanat: Tietovarasto, Liiketoimintatiedon hallinta, ETL

Laurea University of Applied Sciences
Laurea Leppävaara
Business Information Technology

Abstract

Saastamoinen Joni

Data Warehousing and Business Intelligence – a Case Study of TeliaSonera Finland

Year 2012 Pages 51

This thesis focuses on corporations' Datawarehouse (DW) and Business Intelligence (BI) needs and what can be achieved when they are implemented correctly. The author of the thesis will complete his practical training as well as his thesis at TeliaSonera Finland Group IT, Business Intelligence & Data Warehousing unit.

Data Warehousing and Business Intelligence are the fastest evolving areas of information technology. Many companies and government actors have found that the operational systems are not only in themselves sufficient to serve demanding data reporting and analysis needs. Continuous increase in the quantity of information and the recording must be mastered in its entirety in order to obtain important information from processed data - new, old, and pro-active information, to support decision making actors.

One objective was to ascertain if the company's repository PWCIMATICA (PowerCenter Informatica) has the same type of metadata as the company's ETL (extract, transform and load) tool Fileloader produces. It was discerned that in the repository there is very similar metadata as the Fileloader produces. Not all equivalents were found but some cannot be found because some of the metadata is produced by hand.

The file download statistics were found for the most part, as well as the file structure of the source files. In the course of the project more useful metadata were also discovered that the company can use in monitoring loading processes and solving and repairing error situations. A considerable amount of target metadata was also found.

The subject of Data Warehousing and Business Intelligence is highly comprehensive. Consequently, building and maintenance of a DW environment in large companies is a job that needs the whole unit to complete it. It is essential that maintenance teams are in contact with the BI users as early as in the planning stage so that everyone involved in the system gains awareness of it. When everyone is aware of the benefits of the DW and BI environment and is ready to use them, it will definitely show in the company's success.

Keywords: Data Warehouse, Business Intelligence, Metadata

Sisällys

1 Johdanto

Tässä opinnäytetyössä tutustutaan kohdeyrityksen tietovarasto- (Data Warehouse, DW) ja liiketoimintatiedon hallinnan (Business Intelligence, BI) -ympäristöön ja syvemmin paneudutaan varsinaiseen työksiantoon: tarpeeseen löytää repositoriosta teknistä ja operatiivista metadataa yrityksen tarpeisiin. (Repository on varasto, missä säilytetään dataa. Tässä opinnäytetyössä käytetään englanninkielistä termiä, koska se on IT -alalla usein käytetty ja vähiten sekaannusta aiheuttava). Aihealue on harjoittelijalle uusi ja siten mielenkiintoinen sekä haastava kohde lähteä tekemään opinnäytetyötä. Työstä saadusta kokemuksesta on hyötyä jatkossa, sillä DW&BI on tietotekniikka-alalla tärkeä ja hyvin laaja alue.

Yrityksien tietopääoma kasvaa huimaa vauhtia ja sitä varten on kehitetty erilaisia tapoja ottaa tietoa talteen. Tekniikan myötä tallennustekniikat ovat myös kehittyneet huomattavasti ja onkin haastavaa käsitellä isoja tietomassoja tehokkaasti. Haastetta tuo silkka tiedon määrä. Siksi olennaisen tiedon löytäminen ja sen hyödyntäminen on olennainen osa DW&BI -hallintaa. Nykyisin on myös kiinnitetty yhä enemmän huomiota tiedon hyödyntämiseen ja sen jalostamiseen uudeksi tiedoksi. Tieto on yrityksen arvokasta aineetonta pääomaa ja voi oikein hyödynnettynä tuoda huomattavan edun nykypäivän haastavilla markkinoilla.

Opinnäytetyön tekijä suorittaa syventävän työharjoittelun sekä tekee opinnäytetyön TeliaSonera Finlandissa, Group IT, Business Intelligence & Data Warehouse (BI&DW) -yksikössä. Opinnäytetyöaihetta lähdettiin hakemaan kohdeyrityksen kanssa yhdessä ja näin varmistettiin molemminpuolinen hyöty suoritettavasta työstä. Kohdeyrityksellä oli tarjota muutamia aiheita opinnäytetyöksi, jotka käsiteltiin opinnäytetyöprosessin alussa pidetyssä kokouksessa.

Tässä opinnäyteyössä käsitellään ensimmäiseksi aihealueen keskeiset käsitteet. Aluksi kerrotaan miksi DW&BI- ratkaisuja käytetään ja miten niistä voi hyötyä. Tutustutaan myös seuraaviin käsitteisiin: Erilaiset DW:t metadata, ETL-prosessi sekä repositoriot. Viimeiseksi toteutusluvussa esitellään kohdeyrityksen DW&BI- ympäristöä ja selvitetään minkälaista metadataa ETL väline Informatica PowerCenter tuottaa.

TeliaSonera tarjoaa verkkoyhteyksiä ja televiestintäpalveluja yrityksille ja kuluttajille. Palveluja tarjotaan Pohjoismaissa ja Baltian maissa, Euraasian markkinoilla, Venäjällä, Turkissa sekä Espanjassa. TeliaSoneralla on noin 28 tuhatta työntekijää maailmanlaajuisesti ja vuonna 2011 liikevaihto oli noin 12 068 milj. euroa. (TeliaSonera 2012.)

TeliaSonera perustettiin vuonna 2003 kun Telia ja Sonera fuusioituvat. Tänä päivänä yritys toimii viestintäalan edelläkävijänä. Yritys on kehittynyt paikallisista operaattoreista Euroopan

viidenneksi suurimmaksi operaattoriksi noin kahdessakymmenessä vuodessa. (TeliaSonera 2012.)

TeliaSonera Finlandin (jatkossa TSF) Group IT, BI&DW Finland -yksikkö on vastuussa TSF:n BI&DW -toiminnasta. BI&DW -toiminta auttaa yrityksen johtoa ja liiketoimintayksiköitä saamaan paremman ymmärryksen mm. markkinoiden käyttäytymisestä, asiakkaista, prosesseista, tuotteista, henkilökunnasta – koko yrityksen toiminnasta ja suorituskyvystä yleensä. Data voi olla mennyttä, nykyistä tai tulevaisuutta ennakoivaa näkemystä TeliaSoneran liiketoiminta-operaatioista. Tämä auttaa yritystä tehostamaan toimintaansa ja ennakoimaan tulevaisuutta parantaen yrityksen kilpailuetua.

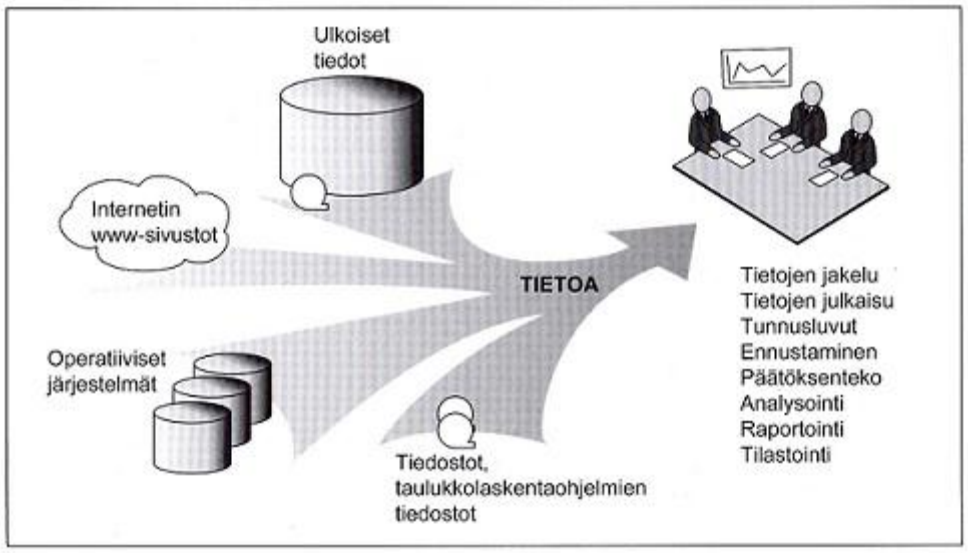
2 Keskeiset käsitteet

2.1 Tietovarastointi ja liiketoimintatiedon hallinta

Tietovarastointi ja liiketoimintatiedon hallinta ovat nopeimmin kehittyviä tietotekniikan alueita. Monet yritykset ja julkishallinnon toimijat ovat huomanneet, että operatiiviset järjestelmät eivät pelkästään itsessään riitä palvelemaan tarpeeksi hyvin vaativia tietojen raportointi- ja analysointitarpeita. Informaation jatkuva määrällinen kasvu ja tallennus tulee hallita kokonaisuudessaan järjestelmällisesti, että saadaan jalostettua tärkeää informaatiota – uutta, vanhaa sekä ennakoivaa, toimijoiden päätöksenteon tueksi. Tietoa jalostetaan omiin tietovarastojen tietokantoihin, mistä sitä haetaan ja hyödynnetään raportointi- ja analysointitarpeisiin. DW- ja BI -ratkaisut lähtevät aina liikkeelle liiketoiminnan tarpeista. (Hovi, Koistinen & Hirvonen 2009, IX.)

Yrityksissä ja julkisissa toimiyhteisöissä on monenlaista pääomaa eli resursseja. Näihin sisältyvät muun muassa. henkilöstö, kiinteistöt, brändit ja tuotantovälineet. Edellä mainituista resursseista pidetään hyvää huolta ja niiden toimintaa sekä tehokkuutta kehitetään jatkuvasti. Myös organisaatioiden tieto on arvokasta pääomaa. Tietojen tuottaminen/syntyminen on vaatinut paljon työtä, laitteistoa ja koulutuksia. Tietoa syntyy ja tallentuu jatkuvalla syötöllä erilaisiin operatiivisiin perusjärjestelmiin. Tällaisia perusjärjestelmiä ovat muun muassa kassakoneohjelmistot ja ERP -sovellukset (Enterprise Resource Planning). Näissä perusjärjestelmissä tiedot ovat yleensä hajallaan, tietorakenteet ovat hankalia ja vaikeasti tulkittavissa. On tärkeää saada tiedot yhtenäisiksi, helposti tulkittaviksi ja saataville, jotta niitä voidaan hyödyntää sujuvasti liiketoiminnan tarpeisiin. (Hovi ym. 2009, XI.)

Organisaatioissa johtajat ja muut toimijat tarvitsevat tietoa päätöksenteon tueksi (katso kuvio 1 alla). Myynnin kehitystä on seurattava useista eri näkökulmista. Edellä mainitut tahot tarvitsevat esimerkiksi: asiakas- ja henkilöstöanalyyseja, mistä ilmenee lukumäärä, koulutus, tuntimäärät, sopimukset jne. Useimmat tiedot ovat tallennettuna yrityksen perusjärjestelmiin, niiden tietokantoihin. Tietojen keräämiseen on sijoitettu merkittäviä summia yrityksen omaisuutta. Yritykset ovat ostaneet erilaisia sovelluksia tai laajoja ERP-järjestelmiä, tai yritys on mahdollisesti tehnyt tai teetättänyt itse sovelluksen, joka vastaa omia tarpeita parhaiten. Myös käyttäjien kouluttaminen eri järjestelmiin vie aikaa ja rahaa. Ajan kuluessa tietokantoihin on varastoitu suuret määrät yritykselle tärkeää liiketoiminnallista tietoa. Tämä tieto muodostaa arvokkaan aineettoman pääoman. (Hovi ym. 2009, 4.)



Kuvio 1: Organisaatiossa tarvitaan monenlaista tietoa (Hovi ym. 2009, 4)

2.2 Tietovarato

Tietovarasto (DW) on yrityksen yhteiskäyttöinen tietokanta. DW:stä puhuttaessa voidaan myös tarkoittaa erilaisia ja kokoisia tietovarastoja yleisesti. Edellä mainittuja tietovarastoja on Datamartti tai EDW (Enterprise Data Warehouse). DW:hen tuodaan tietoa operatiivisista järjestelmistä ja tuotua tietoa voidaan integroida sekä muokata yhteneväiseksi tukemaan BI -käyttäjien tarpeita. Käyttäjät pääsevät BI -työkaluilla omatoimisesti ja suoraan tietoihin käsiksi. ETL- prosessit on usein automatisoitu ajettavaksi öisin, jolloin kuorma on vähäisempi. Automatisoidut ajot vähentävät työtunteja ja virheitä, joita syntyy helposti käsin tehdyissä poiminnoissa. Tiedon integrointi ja muokkausvaihetta kutsutaan ETL- prosessiksi, josta kerrotaan tarkemmin omassa luvussa. DW:ssä voidaan myös säilyttää yrityksen historiatietoja joita voidaan tarpeen mukaan hakea ja tehdä niillä vaikka erilaisia trendianalyysejä (Hovi ym. 2009, 23, 24.)

Tietovarasto on riippumaton liiketoiminnan prosesseista eli niitä ei tarvitse muuttaa tietovarastoa toteutettaessa. Raportoinnista tulee joustavampaa ja tehokkaampaa verrattuna operatiivisten järjestelmien suoraan raportointiin. Eri lähteiden tiedot integroidaan keskitettyyn tietokantaan. Näin kaikki samat aihealueet, esimerkiksi asiakastiedot, saadaan yhteen ja voidaan muodostaa kokonaiskuva asiakkaasta. Tietovaraston avulla osastotason tiedoista saadaan siis yritystason tietoja. Tietojen laatua voidaan helposti seurata, kun tiedot ovat keskitetysti saatavilla. Tietojen laatua seuraamalla ehkäistään väärän informaation kulkeutumista raportteihin. (Hovi ym. 2009, 15,16.)

Johdettuja tietoja ja tunnuslukuja voidaan laskea valmiiksi. Yhteisesti sovituilla kaavoilla lasketut tiedot muodostavat yhden oikean version tiedoista. Näin varmistetaan, että kaikilla on samat tiedot, eikä raportteihin eksy toisistaan eroavia tietoja. Tiedot talletetaan nopeasti ja helposti kyseltävään muotoon. Tietovarastoinnin yhteydessä talletetaan myös metatieto. (Hovi ym. 2009, 15,16.)

Tietovarastoissa säilytetään yrityksen historiaa mahdollistaen paluun vanhoihin tietoihin. Yritys voi siis tarkastella esimerkiksi edellisvuosien tuotteiden, organisaatioiden tilannetta ja hyödyntää niitä tarpeen mukaan. (Hovi ym. 2009, 15,16.)

Tietovarasto vähentää riippuvuutta operatiivisista järjestelmistä. Operatiivisen järjestelmän vaihtaminen käy nopeammin ja sujuvasti kun niiden datat on talletettu keskitettyyn tietovarastoon. Uudesta järjestelmästä rakennetaan vanhaa vastaava rajapinta ETL- prosesseihin ja näin tietojen käyttö voi jatkua saumattomasti. (Hovi ym. 2009, 15,16.)

Operatiivisten järjestelmien raportointia voidaan vähentää ja siten keventää kuormaa, sillä raportointi tehdään tietovarastosta. Operatiivisien järjestelmien suorituskykyä voidaan parantaa poistamalla vanhoja tietoja, joita ei enää tarvita operatiivisella puolella. Vanhan tiedot voidaan poistaa, sillä ne on talletettuna keskitettyyn tietovarastoon. (Hovi ym. 2009, 15,16.)

Organisaation useat erilliset datamartit ilman keskitettyä tietovarastoa synnyttävät monimutkaisia latausketjuja, kun tietoa haetaan ja tuodaan useista datamarteista moneen kertaan. Näiden järjestelmien datavirtakuvaukset ovat hyvin sotkuisia ja vaikeasti ymmärrettävissä. Hyvässä tietovarastoarkkitehtuurissa tiedot poimitaan operatiivista järjestelmistä vain kerran. Tiedot ovat helposti saatavilla. Niitä ei tarvitse enää kaivaa lähdejärjestelmien uumenista. (Hovi ym. 2009, 15,16.)

2.2.1 Tiedon laatu

Tietovaraston (DW) tietoja katsotaan jatkuvasti erilaisista näkökulmista. Tämän mahdollistaa lukuisista operatiivista järjestelmistä yhteen koottu tieto. Operatiivisten järjestelmien tietoja katsotaan vain tietyistä näkökulmista. Yleensä ne on etukäteen ohjelmoituja näkökulmia: sovelluksen näkymä, ikkuna tai raportit. Tämä on riittänyt operatiivisella puolella, mutta DW:n puolella se ei riitä, sillä DW:ssä on tarve analysoida tietoja tarkemmin. Kun tietoja tutkitaan tarkemmin, yleensä paljastuu virheitä ja poikkeavuuksia sekä epätarkkuuksia. Esimerkiksi pankissa tili- ja tapahtumatiedot ovat oikein, mutta asiakkaan ikä, toimiala sekä muut asiakasta luokittelevat tiedot voivat olla puutteellisia. Kun DW:stä halutaan analysoida pankin asiakkaita, haittaa puutteelliset tai väärät tiedot suuresti eikä analysointia välttämättä voida tehdä. (Hovi ym. 2009, 17.)

Mikäli tietojen laatuongelmat eivät paljastu ennen kuin tietovarastoa aletaan testata, on vaarana projektin myöhästyminen. Tietojen laatua pitäisi tarkkailla jo ennen projektin käynnistymistä. Näin puutteet löydetään ajoissa ja niitä voidaan alkaa korjata jo operatiivisessa vaiheessa ennen siirtoa DW:hen. Väärä tieto ei muutu oikeaksi kun se ladataan DW:hen. DW:ssä tietoja ei voi enää muuttaa, sillä ne eivät vastaisi enää operatiivisia järjestelmiä. (Hovi ym. 2009, 68.)

Yrityksen tiedonlaatustrategian laatiminen auttaa yritystä suuresti tiedon laadun varmistamisessa. Tiedonlaatustrategia sisältää tavoitteet tiedonlaadulle. Tavoitteen ei välttämättä tarvitse olla sataprosenttisesti oikeat tiedot, vaan tarkoituksena on määritellä erityyppisille tiedolle omia laatutavoitteita. Erilaiset mittarit auttavat tarkkailemaan tiedon laatutavoitteiden saavuttamista ja auttavat huomaamaan mahdolliset puutteet. (Hovi ym. 2009, 69.)

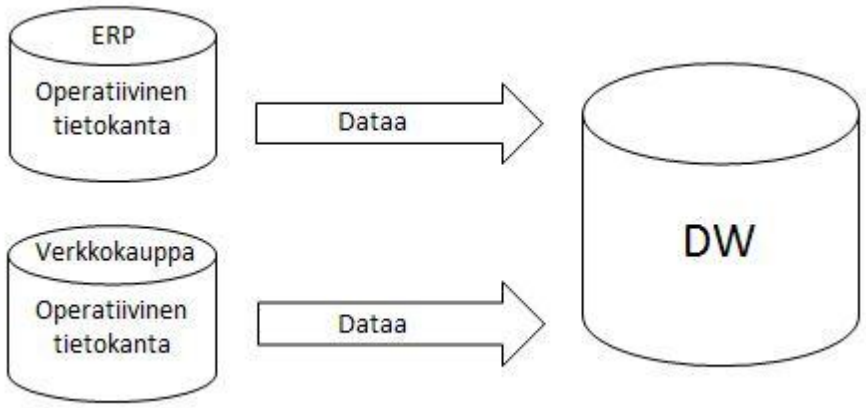
Tiedon laatuun voi vaikuttaa. Hyviä keinoja vaikuttaa tiedonlaatuun on erilaiset menetelmät kuten profilointi, monitorointi sekä yhdistäminen. Profilointi on erilaisten kyselyjen tekemistä, selvitetään tietojen oikeellisuutta. Esimerkiksi voidaan tutkia saman tiedon ilmaisemiseen käytettyjä arvoja. Sukupuoli voidaan ilmaista yhdessä lähdejärjestelmässä käyttämällä arvolla ”N” ja ”M” kun taas toinen järjestelmä käyttää arvoja ”0” ja ”1”, myös muita arvoja voidaan käyttää. Tutkittaessa käytettyjä arvoja, saadaan selville miten monia erilaisia arvoja on käytetty ja nähdään myös tyhjät kentät tai tavallisista poikkeavat arvot esim. ”x”. (Hovi ym. 2009, 69.)

Monitorointi on jatkuvaa tiedon laadun tarkkailua. Eri lähdejärjestelmistä otettuja samaa tarkoittavien tietojen integrointia kutsutaan yhdistämiseksi. Edellä mainittuja menetelmiä voidaan toteuttaa monilla profilointiin, monitorointiin ja yhdistämiseen tarkoitetuilla sovelluksil-

la, joita on esimerkiksi Trillium ja IBM Quality Stage. Myös monet ETL-välineet tarjoavat tiedon laadun tarkkailemiseen käytettäviä sovelluksia. (Hovi ym. 2009, 69.)

2.2.2 Lähdejärjestelmät

Tietovarastoon ladataan dataa useista erilaisista lähteistä. Yleisimmin tietoja poimitaan operatiivisista järjestelmistä tuomalla niiden tuottamia datatiedostoja tai lukemalla tiedot suoraan operatiivisen järjestelmän tietokannoista. Erilaisia lähteitä on itse rakennetut järjestelmät tai kaupalliset sovelluspaketit kuten ERP-svoellukset. Allan oleva kuvio 2 kuvaa tiedon siirtoa lähdejärjestelmistä tietovarastoon (DW). Operatiiviset järjestelmät eivät aina tarjoa kaikkea haluttua tietoa. Esimerkiksi tuotteet halutaan usein järjestää analysointia ja raportteja varten uusiin tuoteryhmiin. Tietovarastoa ei useimmiten haluta päivittää suoraan, vaan se pidetään vain lukukäytössä. Näin voidaan tehdä uudet tuoteryhmäjärjestykset taulukkolaskentaohjelmassa kiinteän muotoisiin taulukoihin, mitkä ETL-prosessin aikana haetaan automatisoidusti yhtenä tietolähteenä mukaan. (Hovi ym. 2009, 18.)



Kuvio 2: Operatiiviset järjestelmät tuottavat dataa ja ne tuodaan keskitettyyn DW:hen.

Ulkoisia tietolähteitä ovat esimerkiksi tilastokeskus, suomenpankki ja posti. Ne ovat hyviä tietolähteitä, kun mietitään tarvetta saada väestö- ja kuntatietoja, valuuttaan liittyvää tietoa yms. Esimerkkinä kuntakohtaisten tietojen hyödyllisyydestä: yritys joka valmistaa vauvanruokapurkkeja voi omista lähdejärjestelmistään selvittää vauvaruoan menekit eri kunnissa. Seuraavaksi yritys pyytää ulkopuolisina tietoina kunnilta vauvojenlukumäärät tai syntyvyysasteen ja vertaa tuloksia omaan myyntiin kunnissa. Mikäli joissain kunnissa on korkea syntyvyysaste ja vähän myyntiä, voi yritys reagoida tähän panostamalla enemmän markkinointiin kyseisessä kunnassa. (Hovi ym. 2009, 18.)

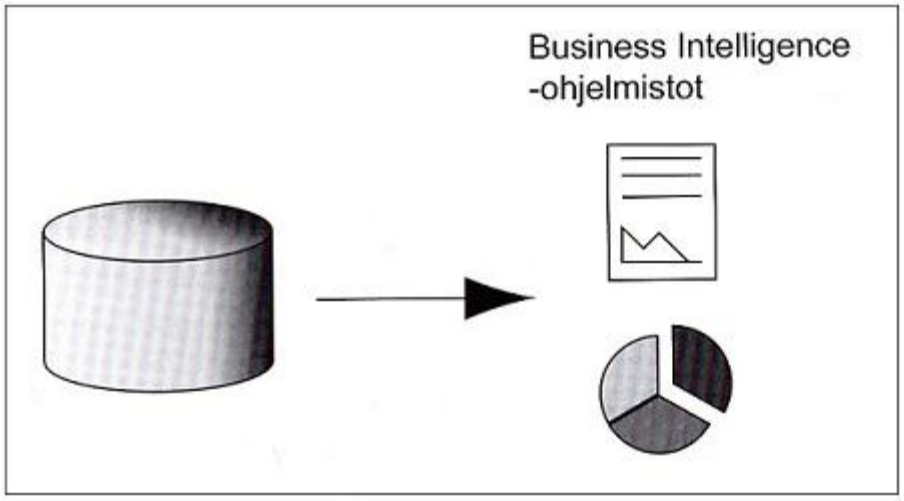
Operatiivinen tietokanta on suunniteltu perussovelluksen käyttöön. Operatiivisiin tietokantoihin syntyy tietoa lähdejärjestelmistä kuten työasemista, automaateista ja kassapäätteistä. Operatiiviset järjestelmät automatisoivat käyttäjien toimintoja ja ne on toteutettu tapahtumankäsittelyjärjestelmiksi (OnLine Transaction Processing System, OLTP). Ne tekevät suuria määriä yhtäaikaisia ja reaaliajassa tapahtumia hyvällä vastausajalla. (Hovi ym. 2009, 22.)

- Laskutusjärjestelmä
- Tilaustenkäsittelyjärjestelmä
- Asiakkuudenhallintajärjestelmä
- Projektinhallinnan sovellus
- Verkkokaupat
- Kassajärjestelmä
- Pankin tilijärjestelmä
- On-line pörssijärjestelmä
- Yrityksen omat sovellukset

Kuvio 3: Yllä on lueteltu erilaisia operatiivisia järjestelmiä.

2.2.3 Datamartti

Datamartti (DataMart, DM) on ns. pienempi tietovarasto. Datamartti on usein suunniteltu aihekohtaiseksi kuten taloushallinnon datamartiksi tai organisaatiokohtaiseksi kuten henkilöstöosaston datamartiksi. Datamartti on tarkoitettu palvelemaan nimenomaan erilaisia kyselyjä ja raportointitarpeita. Se on joko erillinen yhden lähdejärjestelmän tiedot sisältävä tai johdettu, jolloin se on isosta keskitetystä tietovarastosta muodostettu kyselyjä ja raportointia varten tarkoitettu tietovarasto. Datamartit, mitkä sisältävät vain yhden lähdejärjestelmän tietoja, eivät ole varsinaisia tietovarastoja vaan nimenomaan datamartteja. Datamartteja käytetään tukemaan liiketoimintatiedon hallinnan online – tyyppistä käyttöä ja ne on toteutettu ajatellen tietyn käyttäjäryhmän toiveita ja tarpeita. Yleisiä datamartteja on markkinoinnin datamartti tai taloushallinnan datamartti. (Hovi ym. 2009, 24.)



Kuvio 4: Datamartti palvelee kyselyjä ja raportointitarpeita. (Hovi, ym 2009, 24.)

Datamartit voivat olla erillisiä tai yhdenmukaisia. Erilliset datamartit ovat yleensä suppean aihealueen omaavia ja pienissä yrityksissä yleisesti käytetty ratkaisu. Erilliset tai yksittäiset datamartit ovat nopeita toteuttaa ja käyttäjät pääsevät hyötymään niistä nopeasti. Ongelmana niissä on nimenomaan erillisyys. Vaikeuksia tuottaa tehdä kyselyjä, mitkä sisältävät eri aihealueiden tietoja. Tiedot voivat toistua moneen kertaan ja niitä voidaan poimia moneen kertaan, mikä kuormittaa operatiivisia järjestelmiä. Tiedoille ei välttämättä löydy mitään yhteistä linjausta mikä vaikeuttaa yritystason tiedon muodostamista. Tämä on yleensä huonon suunnittelun tulos. Yhdenmukaistetut datamartit ovat Raplh Kimball:n suositus tietovarastolle. Tässä mallissa ei rakenneta yhtä keskitettyä DW:tä vaan monia datamartteja ja ne on nidottu toisiinsa. Tässä ratkaisussa osa dimensioista on yhteisiä ja siksi pitääkin varmistaa, että ne ovat päällekkäisiltä osiltaan yhteneväiset. (Hovi ym. 2009, 25-28.)

2.2.4 EDW

Hyvin yleinen arkkitehtuuri ratkaisu on Enterprise Data Warehouse. Tässä mallissa tiedot kootaan yhteen tai muutamaan isoon tietovarastoon, joita käytetään yhteisesti yrityksessä. Kannoissa on yhdenmukaistettuna monien aihealueiden tietoja, mitä tarkastellaan yritystasolla. EDW:stä voi tehdä eri käyttäjiä varten omia datamartteja, mihin on kohdistettu yhdenkäyttäjäkunnan vaatimat tiedot. Haut tehdään yleensä datamarteista, joten EDW säilyy tiedon yhtenäistämis- ja historiointipaikkana. EDW-ratkaisu voi olla haastava kokonaisuus useine eri tietoaiheineen, mutta on se kuitenkin keskitetty ja yhdenmukaistettu relaatiotietokanta, missä on selkeät roolit käyttäjille. IT-väki tekee EDW:n ja datamartit. BI -käyttäjät yhdistävät kantoihin BI-välineillä ja tekevät raportteja. (Hovi ym. 2009, 27.)

Tietovarastojen ajantasaisuus on tärkeä seikka tietovarastoinnissa. Ennen tiedot päivitettiin noin kuukauden välein, mutta nykyisin tiedot päivitetään kerran päivässä. Välillä voi olla tarpeen tehdä nopeita reaaliaikaisia hakuja (Ad hoc). Tätä varten ovat reaaliaikaiset tietovarastot, joissa voidaan tehdä lähes reaaliaikaisia raportteja ja analysointeja. Raportit eivät ole täysin reaaliaikaisia, vaan yleensä varttitunnin tai tunnin välein tehtävien hakujen tuloksia. Ongelmia tuottaa suorituskyky ja prosessien vaatima aika. Tiedot täytyy saada tietovarastoon välittömästi, eikä ole aikaa tarkistaa tai summata tietoa niin kuin yöllisissä ajoissa. Tästä johtuen tieto on jalostamatonta eli raakaa tietoa. Reaaliaikaiset tietovarastot ovat yleistymässä. (Hovi ym. 2009, 29.)

2.3 Metadata

Metadatan tehtävä on tukea yrityksen liiketoiminnan päätöksentekoa. Yrityksen on tiedettävä ja myös ymmärrettävä, mitä tietoa yrityksen tietovarastossa on. Kaikki tietovaraston tiedot on nimettävä ja määriteltävä ymmärrettävään muotoon. Ilman kunnollisia nimiä ja määrityksiä, on muiden kuin tiedon alkuperäisen tekijän tai käsittelijän vaikea ymmärtää, mistä tiedosta on kyse ja mitä se mahdollisesti ilmaisee. Tästä syystä tiedot ja tietorakenteet on kuvattava ja talletettava kaikkien asianomaisten saataville. Hyvin kuvatut tiedot löytyvät ja pysyvät hallinnassa. (Hovi ym. 2009, 42.) Esimerkkinä edelliseen: tekninen käyttäjä löytää ja ymmärtää epämääräisesti nimetyn ”HX5469” –tiedostossa, sillä hän luultavasti tietää sen sijainnin ennalta tai on tietoinen järjestelmän käyttämästä tavasta nimetä dokumentit. ”HX5469” -tiedostossa on tiedot viimeisen vuosineljänneksen myynnistä. Loppukäyttäjiltä eli BI- käyttäjiltä dokumentti jäisi luultavasti löytymättä, koska sitä ei ole nimetty havainnollistavasti. Näin yrityksen liiketoimintaa tukemaan tehty dokumentti jää hyödyntämättä. Kun tietovaraston tiedot on kuvattu ja nimetty hyvin ja ymmärrettävästi, antaa se loppukäyttäjille myös motivaatiota hyödyntää olemassa olevia tietoja. Tieto täyttää tarkoituksensa vain kun se on saatavilla, ymmärrettävää ja käytettävissä.

Metadatan yleinen määritelmä on ”tietoa tiedoista”. Metatiedot ovat kuvauksia tietovaraston eri objekteista, kuten liiketoiminnan alueista, tauluista, tiedoista ja näkymistä. Käyttäjä voi tietoa tarvitessaan tutkia metatiedoista, onko kyseistä tietoa saatavilla. (Hovi ym. 2009, 43.)

David Marco on määritellyt metadata käsitteen tavalla, joka antaa hyvän käsityksen metadatan laajuudesta ja siitä mitä kaikkea tietoa tiedosta voi olla: Metadataa on kaikki fyysinen ja ei-fyysinen tieto, joka liittyy ohjelmiin ja muuhun mediaan, henkilökunnan osaamiseen, yrityksien sisäiseen ja ulkoiseen tietoon. Tietoihin kuuluu myös kaikki informaatio liiketoiminta prosesseista, tiedon säännöistä ja sidoksista sekä tiedon rakenne ja sijainti (2000, 5).

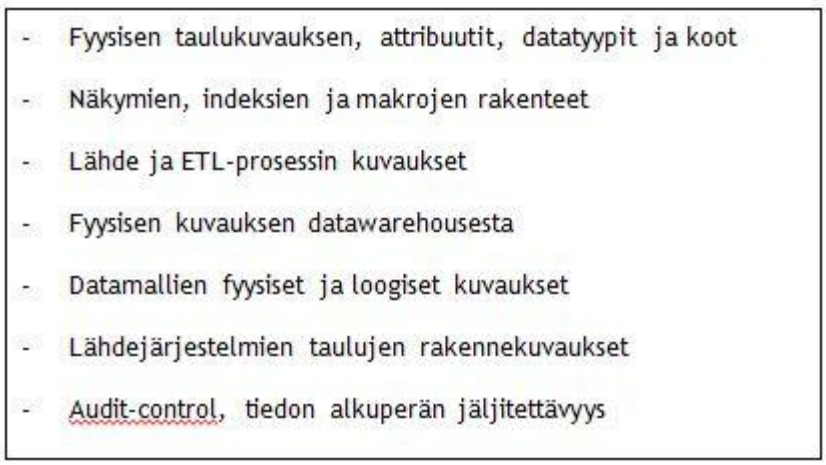
Metadata on usein luokiteltu kolmeen ryhmään: Tekninen metadata, liiketoiminta metadata ja operatiivinen metadata. Seuraavissa kappaleissa käsitellään yleiset metadataluokat ja annetaan esimerkkejä niiden sisältämästä metadatasta. Koska käsite metadata on laaja, käytetään metadatasta puhuttaessa myös muita kuin luokkanimiä. Usein käytetään tarkentavia aihekohtaisia nimiä, kuten target metadata, mikä viittaa tietovaraston tai datamartin kohdetauluihin.

2.3.1 Tekninen metadata

Tekninen metadata kertoo tarkan rakennekuvauksen tietovarastosta tai tietokannasta sekä auttaa ennakoimaan järjestelmiin tehtävien muutoksien vaikutusta kokonaisuuteen. Se myös tukee tarkkaa dokumentointia ja tietovarastossa olevien tietojen tulkintaa. Teknistä metadataa löytyy tyypillisesti ohjelmista, tietokannan järjestelmä tauluista tai ETL -työkaluista ja kuvaa tietoa teknisestä näkökulmasta. (TS Metadata Requirements R001, 7.)

Tekninen metadata auttaa kehittäjiä ja ohjelmoijia antamalla informaatiota päätöksenteko järjestelmistä ja operationaalisista järjestelmistä, mitä he tarvitsevat ylläpitääkseen ja kehittääkseen näitä järjestelmiä. Yrityksellä voi olla tarve uudelleen määritellä maantieteelliset myyntialueet. Teknisen metadatan avulla voidaan listata kaikki ohjelmat, taulut ja järjestelmät, mitkä sisältävät maantieteellistä myyntitietoa. Tämän tiedon avulla kehittäjät voivat arvioida nopeasti tehtävää varten tarvittavan työmäärän ja resurssien tarpeen. Se myös paljastaa muut järjestelmät, mihin muutos voi vaikuttaa. Ilman teknistä metadataa eri prosessien analysointi ja toteutus vaikeutuu ja vie huomattavasti enemmän aikaa. (Marco 2000, 49.)

Kuvio 5: Teknistä metadataa on monenlaista. Se ilmaisee:



2.3.2 Liiketoiminta metadata

Liiketoiminta metadata (BI -metadata) toimittaa loppukäyttäjille helposti ymmärrettävät näkymät tietovaraston sisällöstä. BI -käyttäjät vaativat yhä enemmän semanttisempaa sisältöä tiedosta mitä on varastoitu tietovarastoon. He myös haluavat etsiä ja löytää tietoa BI -käyttäjille tutuilla termeillä. Tietovaraston käyttäjät yleensä hakevat BI -metadataa. (TS Metadata Requirements R001, 2012.)

Liiketoiminta metadata kuvaa tietoa liiketoimintatermein, kuten nimeämällä attribuutit ymmärrettävään muotoon tai antamalla attribuuteille ja kaavoille havainnollistavat kuvaukset. Kuvaileva BI -metadatan tuottaminen vaatii enemmän järjestelmällisyyttä ja tarkkuutta, sillä ohjelmat eivät tuota business metadataa automaattisesti, eivätkä järjestelmät vaadi sitä toimiakseen. Business metadata muodostuu tiedon tuottajien toimesta heidän antaessa selvät ja tarkat kuvakset eri tiedoille. Raportin tuottajat tarvitsevat liiketoimintakuvauksia tuottaessaan tarkkoja raportteja liiketoiminnasta. Hyvin kuvatut tiedot auttavat ymmärtämään tiedon sisällön ja merkityksen helpommin. Se myös kertoo mitä tietoa sinulla on, mistä se on tullut, mitä se tarkoittaa ja mikä sen suhde on muuhun tietovaraston tietoon. Huonosti hoidetussa tietovarastossa tietoja ei ole kuvattu lainkaan tai todella heikosti. Tiedon omistajan tulee olla tarkkana miettiessään sopivaa ja oikeaa kuvausta tiedoilleen. BI -metadata on avain kunnolliseen tietovarasto ja liiketoimintatiedon -dokumentaatioon. (TS Metadata Requirements R001, 2012.)

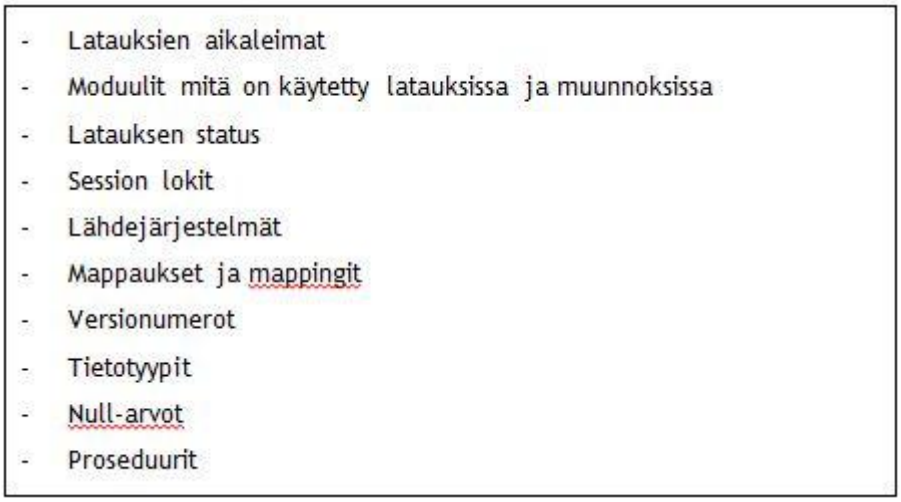
Kuvio 6: Alla esimerkkejä business metadatasta

- Johdonmukaiset liiketoimintatermit ja kuvaukset tauluille sekä attribuuteille
- Tietoelementtien liiketoimintakuvaukset
- Liiketoiminnan säännöt ja oikeat arvot (rules and values)
- Loogiset datahaarat, muunnokset ja mappaukset, mitä käytetään tukemaan analyyttistä raportointia ja tietojenmallinnusta
- Kohdealueen kuvaukset
- Tiedon alkuperän
- Loogisen datamallin
- Päätöksenteko järjestelmän päivitys päivämäärät
- Tiedon omistaja
- Datan laadun statistiikat

2.3.3 Operatiivinen metadata

Operatiivinen metadata kaappaa informaatiota prosesseista, mitkä vaikuttavat dataan. Tällaisia prosesseja on muun muassa lähdetiedostojen lataukset, varmuuskopioinnit ja statistiikat puuttuvasta tiedosta. Operatiivinen metadata ylläpitää tietoa lähteistä, muutoksista ja tietovaraston käytöstä. Näin varmistetaan, että tietovaraston tiedot pysyvät luotettavina ja eheinä. ETL-prosessissa kaikki avaintieto ajoista talletetaan. Avaintietoa on muun muassa aloitusaika, lopetusaika, ladatut tai hylätyt rivit ja virhelokit. Edellä mainitut avaintiedot auttavat latausprosessien seurannassa ja mahdollisten virhetilanteiden korjauksessa. Operatiivinen metadata palvelee eniten tietovaraston tuki- ja ylläpitohenkilöitä. (TS Metadata Requirements R001, 2012.)

Kuvio 7: Operatiivinen metadata ilmaisee seuraavia asioita:



2.3.4 Metadatamalli

Metadatan tallentaminen standardin metamallin mukaisesti on tärkeää, sillä se ratkaisee monia IT-haasteita, joita esiintyy nykypäivän liiketoiminnassa. Eri mallien mukaan rakennetut repositoriot ovat hankalia tilanteessa, missä on tarvetta yhdistää useampi repositorio. Tämmöinen tilanne on esimerkiksi kahden yrityksen fuusioituminen. Repositorioiden yhdistämisellä halutaan käyttöön molempien yrityksien arvokas tietopääoma, mitä on esimerkiksi asiakastiedot.

Eri metadatamallin mukaiset repositoriot on hankalaa ja kallista yhdistää, sillä yrityksien ohjelmat eivät välttämättä tue samoja rajapintoja. Monet yritykset ovat ostaneet monia eri toimittajien ohjelmia, joista jokainen tukee erilaista metadatamallia. Näille ohjelmille on rakennettava omat rajapintansa, mikä on kallista ja aikaa vievää. Samaa metadatamallia tukevat ohjelmistot ja repositoriot on helpompi yhdistää ja uusien kokonaisuuksien integrointi nykyisiin järjestelmiin sujuu vaivattomammin. Kansainväliset yritykset pystyvät helposti ja nopeasti jakamaan dataa eri maiden yksikköjen välillä, kun on noudatettu yhtenäistä metadatamallia.

Ennen oli kaksi vakiintunutta metadatamallia: Open Information Model (OIM) ja Common Warehouse Model (CWM). OIM on MetaData Coalitionin (MDC) perustama standardi. OIM on tarkoitettu työkaluksi tukemaan eri teknologioiden ja yritysten tiedon yhteensopivuutta hyödyntämällä jaettua tietomallia. OIM kattaa osa-alueet tietojärjestelmien kehittämisestä käyttöönottoon. OIM käyttää UML:n konsepteja ns.alimalleissa (sub-model), mitä käytetään kuvaamaan metadataa.

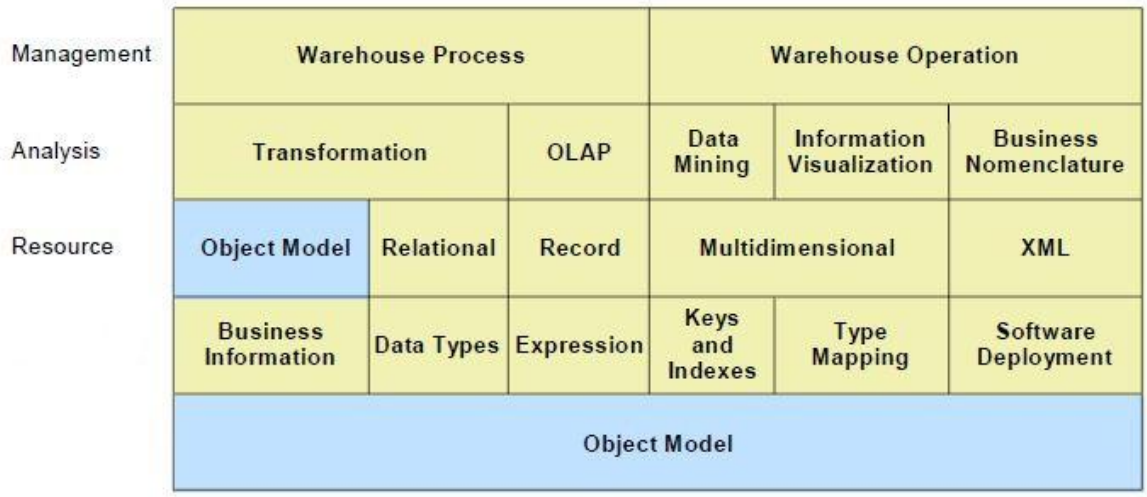
Common Warehouse Model on Object Managment Group:n (OMG) standardi. CWM on tehty helpottamaan metadatan siirtoa eri tietovarastotyökalujen ja repositorioiden välillä. CWM perustuu seuraaviin OMG standardeihin: UML, MOF, metamallinnuksen ja metadata repository-standardi, XMI, XML. OIM ja OMG –standardit yhdistyivät tarkoituksena luoda yhteinen standardi. OIM -standardista tuli osa CWM -kokonaisuutta. (Vetterli, 1999.)

Metadatamalli-standardi

CWM – standardi määrittelee rajapintoja. Rajapintojen avulla voidaan siirtää tietovarastointi ja liiketoimintatiedon hallinnan -metadataa eri tietovarasto työkalujen ja repositorioiden välillä. CWM -metamalli sisältää monia alimalleja (katso kuvio 8 alla), mitkä kuvaavat Common warehouse metadataa:

- Data resurssit: Nämä sisältävät metamallit, jotka kuvaavat oliot, relaatiot, tallenteet, dimensiot, XML-datalähteet.
- Data analysointi: Nämä sisältävät metamallit, jotka kuvaavat tiedon muunnokset, OLAP, Data Mining, tiedon kuvakset, liiketoimintatermit.
- Datan hallinta: Nämä sisältävät metamallit, jotka kuvaavat tietovaraston prosessit ja operaatioiden tulokset.

(Common Warehouse Metamodel, v1.1, 2003.)



Kuvio 8: CWM metamalli on suunniteltu hyödyntämään Object Model:n käyttöä (Object Model on yksi UML:n alimalli). (Common Warehouse Metamodel, v1.1, 2003)

CWM käyttää seuraavia standardeja:

- UML – Unified Modelling Lanaguge. OMG –mallinnus-standardi
- MOF – Meta Object Facility. OMG -metamallinnus ja metadata repository -standardi
- XMI – XML Metadata Interchance. OMG metadata siirto-standardi

(Common Warehouse Metamodel, v1.1, 2003.)

2.3.5 Metadata repositorio

Metadata repository on hieno nimi tietovarastolle, joka on suunniteltu keräämään, säilyttämään ja levittämään metadataa. Repositorio on vastuussa metadatan luokittelemisesta ja jatkuvan fyysisen metadatan pysyvästä varastoinnista.

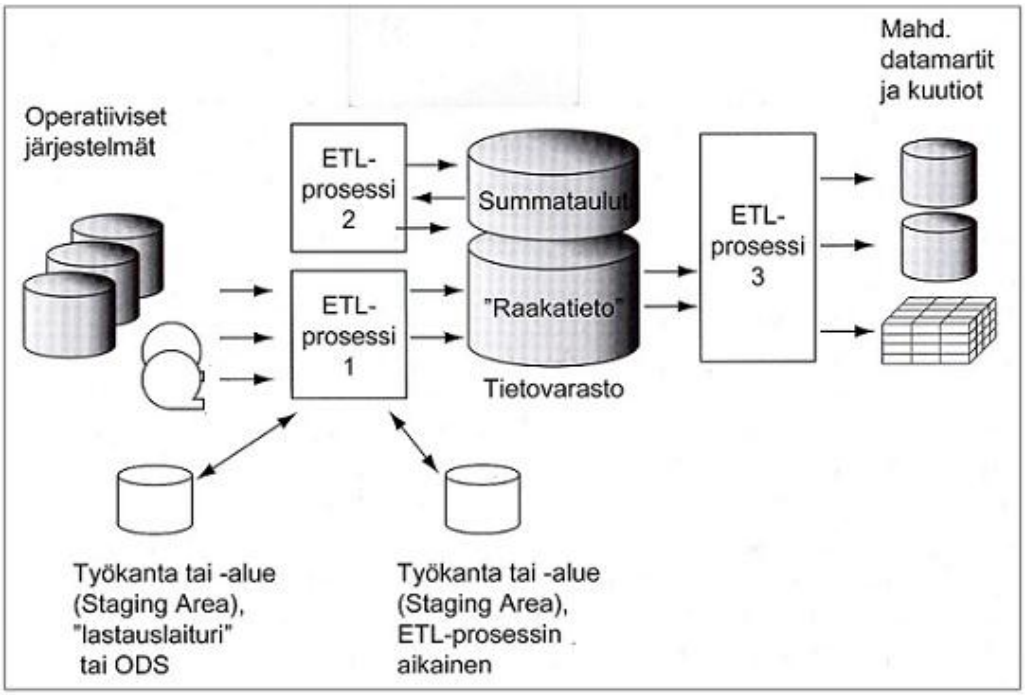
Metadata repository:n tulisi olla geneerinen, kokonaisvaltainen, ajantasainen ja historiallinen metadatavarasto. Repositorion geneerisyys tarkoittaa sitä, että fyysinen metamalli tallentaa metadatan sen aihealueen mukaan repositorioon, ei sovelluksen tavan mukaan. Esimerkiksi geneerinen metamalli sisältää attribuutin nimeltä ”DATABASE_PHYS_NAME”, joka sisältää kaikki tietokannan fyysiset tietokantanimet yrityksen sisällä. Sovelluksen metamallin mukaan attribuutit nimettäisiin sovelluksen nimellä, esimerkiksi ”ORACLE_PHYS_NAME”. Sovelluksen omat metamallit ovat ongelma, sillä metadatan aihealueet muuttuvat. Yritys voi haluta tulevaisuudessa siirtyä käyttämään erilaista tietokantaa yhteensopivuussyistä. Standardin metamallin mukaan nimetyt attribuutit ovat käteviä, sillä niitä ei tarvitse muuttaa. (Marco 2000, 36.)

Metadata repositorio tarjoaa kokonaisvaltaisen näkymän yrityksen metadatan aihealueisiin. Käyttäjän pitäisi olla mahdollista nähdä kaikki kokonaisuudet yrityksestä, eikä vain esimerkiksi Oracleen ladattuja kokonaisuuksia. Repositorio sisältää myös nykyistä ja tulevaa metadataa tarkoittaen, että metadata päivitetään tietyin väliajoin vastaamaan nykyisiä ja tulevaisuuden teknisiä sekä liiketoiminta-ympäristöjä. Kun metadata repositoriota päivitetään jatkuvasti, se pysyy arvokkaana ja tärkeänä yrityksen toiminnalle. Metadata repositoriossa on syytä säilyttää myös historiatiedot. Hyvässä repositoriossa säilytetään metadatan vanhoja näkymiä vaikka tieto ja metadata vaatimukset muuttuvat ajan kuluessa. Tämä auttaa yrityksiä ymmärtämään, miten liiketoiminta on kehittynyt aikaa myöten yrityksen kasvaessa. (Marco 2000, 36-37.)

2.4 ETL

ETL-prosessissa tiedot luetaan operatiivisista järjestelmistä tai siirtotiedoista, muokataan tietovarastotietokannan muotoon ja talletetaan DW:hen. ETL tulee sanoista Extract, Transform, Load, eli poiminta, muokkaus ja lataus. Poiminnassa valitaan lähdejärjestelmistä haluttavat tiedot ja tallennetaan ne tiedostoon. Muokkauksen aikana tarkistetaan virheet, historiointi määritykset, koodien muunnokset ja muunnetaan ne oman datamallin mukaiseksi. Lopuksi muutetut tiedot ladataan DW:n tauluihin. Tiedot jalostetaan ajon aikana valmiiksi, näin päiväsaikaan tehtävät kyselyt on nopeasti tehtävissä.

ETL toteutus on tietokantapohjaista eräajosovelluksen ohjelmointia. Tätä varten on kehitetty ETL -työkalut. Työkalut sisältävät yleensä hyvän ja selkeän käyttöliittymän, jolla on sujuvaa tehdä työtä. BI -välineillä tehtävää tietojen raportointia ja analysointia voidaan tehostaa käyttämällä aikaa tietojen integrointiin, muokkaukseen ja jalostukseen, muodostaen näin helppokäyttöisiä rakenteita. (Hovi ym. 2009, 48.)



Kuvio 9: ETL-prosessi (Hovi ym. 2009, 49)

Tietovaraston ETL -prosessit ovat pääpiirteittäin samantapaisia, mutta usein niitä on muokattu oman ympäristön tarpeita varten. Yllä olevassa kuviossa 9 on kuvattu tyypillisen tietovarastoympäristön ETL -prosessin vaiheet. ETL -prosessi 1 kuvaa raakatietojen poimintaa operatiivisista järjestelmistä sekä muokkaamista ja lataamista tietovarastoon. Operatiivisista järjestelmistä tuodaan viimeisimmät tapahtumat staging arean I-tauluihin (Input). Staging area on

ETL -prosessissa käytetty työtietokanta, missä tiedot muokataan valmiiksi ja odottavat siirtoa tietovarastoon. Staging area on suljettu käyttäjiltä ja se voi sijaita fyysisesti samalla tai eri palvelimella kuin tietovarasto. I-taulut muokataan, tarkistetaan ja muunnetaan käyttäen apuna W-tauluja (Work). Valmiit tiedot kirjoitetaan O-tauluihin (Output) ja ladataan tietovarastoon. (Hovi ym. 2009, 48-49.)

ETL –prosessin toisessa vaiheessa muodostetaan summataulut ladatuista tiedoista. Summatauluja ovat esimerkiksi päiväkohtaisten myyntien koosteita kuukausikohtaisiksi. Summataulut nopeuttavat hakuja tietovarastosta, sillä ne ovat valmiiksi laskettuja kokonaisuuksia. ETL-prosessin kolmannessa vaiheessa muodostetaan datamartit ja kuutiot raportointia ja tutkimista varten. ETL -prosessien tulisi muodostaa saumaton kokonaisuus, joka ajetaan edellä mainitussa järjestyksessä yleensä öisin, jolloin kuorma on vähäisempi. (Hovi ym. 2009, 48-49.)

2.4.1 Menetelmät

ETL-toteutukseen on kaksi perusmenetelmää: Työntö- ja vetomenetelmä. Työntömenetelmässä poimitaan halutut tiedot ja tallennetaan ne tiedostoihin operatiivisessa järjestelmässä. Tiedostot siirretään tietovaraston lastauslaiturille, muokataan ja lopuksi luetaan tietovaraston tauluihin.

Työntöperiaatteen etuja ovat: tiedot luetaan operatiivisten järjestelmien puolella ja ne ajastetaan niin, että niitä ei lueta kesken päivityseräajojen. Operatiivisten järjestelmien rakenteet ovat usein hyvin monimutkaisia, siksi poiminnan suorittaa operatiivisen järjestelmän asiantuntija, kenelle rakenteet ovat tuttuja. Latausajoihin syntyy myös hyvä rajapinta, mistä on muun muassa seuraava etu: latausajot voidaan uusia nopeammin, kun ei tarvitse palata operatiiviseen järjestelmään hakemaan tietoja, jotka voivat olla kaiken lisäksi muuttuneet ja näin tilanne olisi menetetty.

Operatiivisten järjestelmien vaihtaminen on mahdollista. Täytyy vain tilata uudelta järjestelmältä samantyyppiset tiedostot ja tietovaraston toiminta voi jatkua keskeytyksettä. Ongelma työntömenetelmässä on, että ETL -välineillä tiedostoista lukeminen on hankalampaa kuin suoraan operatiivisesta kannasta lukeminen, sillä tietokuvauksia ei saa helposti mukaan. Menettelytavassa on myös enemmän vaiheita, joten kustannukset voivat nousta niiden mukana. (Hovi ym. 2009, 51.)

Vetomenetelmässä tiedot luetaan suoraan operatiivisesta järjestelmästä ETL -välineellä. Kun ETL -välineellä kytkeydytään suoraan operatiiviseen kantaan, saadaan tietojen nimet ja kuvaukset mukaan ETL -välineeseen. Ajastetuissa latausajoissa ETL -välineet lukevat suoraan operatiivisten järjestelmien tietokantoja. Tämän menetelmän etuja on: Yksinkertainen toteutus

(ei välitiedostoja), mikäli kannat ovat tuttuja se antaa joustavuutta ja nopeutta työskentelyyn. Ongelmina on selkeän rajapinnan uupuminen, operatiivisen järjestelmän kanta ei ole tuttu, joka johtaa kalliiden väärinymmärryksien syntymiseen. Tiedostot voidaan lukea kesken kannan päivityksen, jolloin voidaan saada eri ajankohdilta tietoa samaan pakettiin. Paketti ei siis ole enää yksi tilanne. Virhetilanteiden uusinnat voivat osoittautua hankaliksi. Mikäli tilanne on jo kerinnyt muuttua, uusintaa ei voida tehdä eli tilanne on menetetty. Mikäli operatiivinen järjestelmä vaihtuu, joudutaan kirjoittamaan koko ETL -prosessi uudestaan. (Hovi ym. 2009, 52.)

2.4.2 ETL -toteutus

Toteutus tapahtuu ohjelmoimalla tai käyttämällä ETL -välineitä. Ohjelmoiden toteutettaessa käytetään relaatiokannan valmiita proseduureja. Proseduureja voidaan tehdä mm. Javalla, MS SQL Serverillä, Oraclen PL/SQL-kielellä. Ohjelmoiden toteutus teettää enemmän töitä ja muodostaa helposti henkilöriippuvaisuuksia, sillä monet ohjelmoivat omalla tavallaan, mitä muiden on hankala kehittää ja ylläpitää. ETL -välineillä tehtäessä syntyy myös koodia, mutta se on visuaalisemmin käsiteltävissä. Erilaiset ETL -välineet ovatkin ns.sovelluskehittimiä. Ne tarjoavat valmiita ratkaisuja, jotka nopeuttavat työntekoa. ETL -välineet toimivat tietyllä toimintaperiaatteella ja ovat yleensä graafisia sovelluksia. Tämä vähentää henkilöriippuvuuksia. Toimintaperiaatteen ollessa tuttu, ymmärtävät uudet käyttäjät jo tehtyjä ratkaisuja. (Hovi ym. 2009, 53-61.)

ETL -työkalun ominaisuuksia:

- Graafinen suunnitteluosio
- Audit trail
- Laadun profilointi
- Latausajojen lokit ja tiedostot
- FTP tuki
- Mappaus eli tietojen yhdistely
- Tuplarivien poisto
- SQL-käskyjen ja tietokantaproseduurien ajo
- Tuki monenlaisille tietolähteille
- Latausprosessien versionhallinta

(Hovi ym. 2009, 61.)

3 Liiketoimintatiedon hallinta

Liiketoimintatiedon hallinta (Business Intelligence, BI) on tietojen muokkausta entistä informatiivisemmaksi. Se auttaa tekemään oikeita ratkaisuja kilpailutoiminnan kiristyessä. Yrityksiä kiinnostaa ymmärtää oman yrityksen tämänhetkinen tilanne sekä miten siihen on tultu. Myös mahdolliset ennusteet tulevaisuuden markkinoista analysoimalla maailman kehitystä ja edellisiä vuosia, ovat suuri etu jokaiselle yritykselle. Päätöksiä on helpompi tehdä, kun on saatavilla luotettavaa sekä ajan tasalla olevaa informaatiota.
(Hovi ym. 2009, 73–79.)

Edellä mainittua toimintaa varten on kehitetty BI -ratkaisuja, joilla päästään käsiksi toimintaa kuvaavaan informaatioon. BI -toimintaa on esimerkiksi raporttien- luominen, päivittäminen ja analysointi. Loppukäyttäjien ei tarvitse lukea tietokantoja informaatiota saadakseen, vaan ne ladataan erilaisiin BI –sovelluksiin. Niissä tiedot on esitetty visuaalisesti ja helposti ymmärrettävästi. Sovellukset ovat usein point-and-click-ratkaisuja, joiden käyttämistä voisi verrata netissä surffaamiseen. Yritykset ovat aina keränneet tietoja ja nykyisin talteenottokin on nopeaa ja helposti tehtävissä, sillä levykapasiteetit ja tallennusmenetelmät ovat kehittyneet. Ongelma on siinä, että miten tietoa voidaan hyödyntää tehokkaasti. Miten tiedot jalostetaan enemmän informatiivisiksi, miten niitä tulisi analysoida. Tiedot on saatava hyödynnettyä sujuvasti ja nopeasti. (Hovi ym. 2009, 73–79.)

Tietoa syntyy operatiivisissa järjestelmissä jatkuvasti. Tietoa syntyy myös itsestään muiden toimien ohella. Esimerkiksi kännykän kanssa liikkuminen tuottaa jatkuvasti tietoa käyttäjän tekemisistä ja sijainnista. Muuta digitaalista tiedon keräämistä on esimerkiksi kaupankassojen kuittien lukeminen. Ostoksia seuraamalla saadaan selville mitä tuotetta myydään tällä hetkellä paljon ja voidaan analysoida onko kyseisen tuotteen markkinointi vaikuttanut myyntiin. Kauppojen kanta-asiakaskortit mahdollistavat asiakkaiden kulutustottumuksien seurannan, mitä voidaan verrata vaikka asiakkaan asuinalueeseen. Minkälaisille tuotteille on kysyntää esimerkiksi pientaloalueella tai keskustan tuntumassa. Kanta-asiakaskorttien avulla saadaan ostoksiin liitettyä myös ikä ja sukupuoli sekä jo mainittu asuinalue. Eli asiakkaasta saadaan kohtuullisen hyvä kokonaiskuva ja näin markkinointia sekä tuotetarjontaa voidaan kohdistaa tehokkaasti. (Hovi ym. 2009, 73–79.)

Nykyisin yritysten toimintatahti on kiihtynyt ja tulee kiihtymään entisestään. Tämä tarkoittaa, että informaatiota on hyödynnettävä entistä nopeammin päätöksiä varten. Tavanomaisimmin raportteja tuotetaan kuukausi tai päivätasolla, mutta kun on kriisitilanne, esimerkiksi lama, niin voi olla tarpeen tuottaa raportteja useasti päivässä. Kaikkea tietoa ei välttämättä tarvitse saada usein. Yleensä onkin päätetty ennalta tietyt avaintunnusluvut (Key Performance Indicators, KPI) mitä seurataan tarkkaan ja usein. KPI voi olla tietyn myymälän päivän myynti tai yrityksen omavaraisuusaste. Tiedon keräämiselle on myös lain määräämiä velvoitteita. On säädetty laki, joka määrää yritykset tarvittaessa todentamaan käytetyn tiedot alkuperän (Audit Trail). Esimerkiksi pörssiyhtiöiden ja pankkien kohdalla tämä on tärkeää. (Hovi ym. 2009, 73–79.)

Liiketoimintatiedon hallinnan tulkinta

Liiketoimintatiedon hallinnalla voidaan tarkoittaa myös yrityksen sisäistä tai ulkoista näkemystä. Sisäinen näkemys on kvantitatiivinen näkemys. Tällä tarkoitetaan yrityksen sisäisesti keräämään datan analysointia ja hallintaa. Lähteinä toimivat yrityksen tietokannat ja järjestelmät. Tieto on relaatiokantoihin tallennettavaa ja hyvin usein numeerista dataa (myyntiluvut) eli strukturoitua dataa. (Hovi ym. 2009, 80–83.)

Ulkoisesta näkemyksestä puhuttaessa tarkoitetaan kvalitatiivista näkemystä. Kvalitatiivinen näkemys on kilpailijoista ja markkinoinnista saatavan datan hyödyntämistä. Dataa saadaan usein julkisista lähteistä, kuten uutisista tai tilastokeskuksesta. Data on strukturoimatonta eli dokumenttipohjaista. (Hovi ym. 2009, 80–83.)

Liiketoimintatiedon hallinnan tavoitteita:

- Nopeuttaa ja parantaa organisaatioiden kykyä tehdä päätöksiä
- Vastata käyttäjien tietotarpeisiin oikea-aikaisesti
- Tukea organisaation strategiaa ja tavoitteisiin pääsyä
- Parantaa käyttäjien omatoimisuutta tietotarpeiden suhteen
- Vähentää kustannuksia ja parantaa operatiivista tehokkuutta
- Käytettävät ratkaisut lähtevät liiketoiminnan tarpeista (järjestelmät rakennetaan BI:n tarpeisiin, ei toisinpäin).
- Mahdollistaa faktajohtaminen

(Hovi ym. 2009, 80–83.)

3.1.1 Data Mining

Data mining -menetelmällä tarkoitetaan tiedon louhimista. Menetelmässä pyritään löytämään jo olemassa olevasta informaatiosta uutta tietoa tarkastelemalla tietoa useista eri näkökulmista. Tietovarastosta voidaan etsiä tietoja yhdistäviä tekijöitä, jotka helpottavat liiketoiminnan ennakoimista. Menetelmät ovat yleensä puoliautomaattisia matemaattisia ja tilastollisia analysointimenetelmiä. Data mining analyysien tuottaminen pelkästään perinteisillä raporteilla veisi huomattavan paljon aikaa ja tulisi kalliiksi. (Hovi ym. 2009, 99)

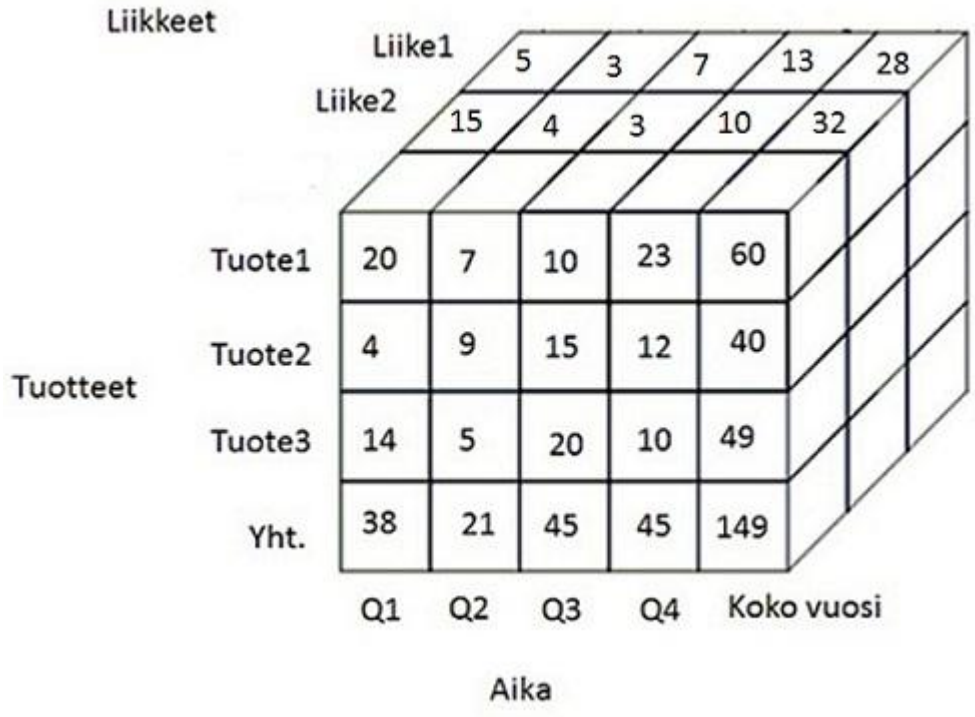
Esimerkiksi ostokäyttäytymis-dataa voidaan analysoida data mining menetelmällä. Tutkitaan ostoskorien sisältöjä ja etsitään useita toistuvia samankaltaisuuksia. Kun huomataan tiettyjen tuotteiden esiintyvän usein samassa ostoskorissa, voidaan ne tuotteet sijoittaa kaupassa lähelle toisiaan vastaamaan toistensa kanssa hyvin meneviä ostoksia. Esimerkkinä ostokäyttäytymisen tutkimisesta: eräs liike havaitsi, että miehet ostavat vaippoja torstaisin ja lauantaisin. He myös yleensä ostivat olutta samalla. Tarkempi tutkimus kertoi heidän tekevän suuret ostokset yleensä lauantaisin ja torstaisin hankittiin vain muutamia tuotteita. Tästä pääteltiin, että miehet ostivat oluet valmiiksi viikonloppua varten. Tämän seurauksena myymälät pystyivät varautumaan tilanteeseen varmistamalla, että vaippojen lähellä näkyi olutmainoksia ja oluet myydään täyteen hintaa torstaisin, ei alennuksella. Tarkoituksena voi olla myös tavoite aiheuttaa heräteostoksia. Myös tyypillinen esimerkki data mining analyysista on verkkokaupat, joissa tuotteen ostajalle tarjotaan muita samankaltaisia tuotteita tai tuotteita, mitä muut asiakkaat ovat ostaneet samalla kertaa. (Palace. 1996; Hovi ym. 2009, 99)

Yksityisyys voi olla ongelma data mining analyyseissä. Analyysien avulla voidaan koostaa niin tarkka kuva yhdestä henkilöstä, että se rikkoo jo henkilön yksityisyyttä. Menetelmän avulla saadaan selville henkilön tottumukset eli mitä hän ostaa, syö, missä hän suorittaa ostokset. Erilaisten etukorttien tietojen avulla saadaan vielä osoitetiedot ja syntymäajat yhdistettyä yksilöön. Tietoja käyttävien onkin varmistettava tietojen oikeanlainen käyttö organisaatiossa. Tietoja tulee käsitellä anonyymisti eli ei tutkita tietyn henkilön tekemisiä, tottumuksia tai asuinsijaintia. Analysoitavat tiedot käsitellään joukoissa esim. ikäryhmien ostoskäyttäytymiset tai asuinalueen kulutus. (Palace. 1996)

Tietojen yhdistäminen voi joissain tapauksissa olla ongelma. Toisinaan yrityksissä (pankit) tieto voi olla jaoteltu turvallisuussyistä eri tietokantoihin ja palvelimille. Muun muassa asiakkaan luottokortin tiedot voivat olla eri sijainneissa. Ohjelmien pitää siirtää tietoa useista eri sijainneista kootakseen kokonaiskuvaa, mikä rasittaa tietojärjestelmiä. (Palace. 1996)

3.1.2 OLAP

On-Line Analytical Processing on tekniikka, jolla voi tarkastella tietoa moniulotteisesti analysointia varten. Dataa voidaan kuvata OLAP -kuutioilla. Kuutio on tietovaraston tiedoista valmiiksi koostettu strateginen ja moniulotteinen näkymä. Kuutio kootaan yhdistämällä dimensioita (ulottuvuus). Dimensioita voivat olla aika, tuote, sijainti jne. Näin muodostuu kolmiulotteinen kuutio, jonka jokainen sivu kuvaa yhtä dimensiota. Jokaiseen dimensioon kuuluu joukko arvoja esim. summattu luku.



Kuvio 10: OLAP-kuutio

Yllä olevassa kuviossa 10 on yksinkertainen esimerkki OLAP-kuutiosta. Kuutiossa on 3 dimensiota: liikkeet, tuotteet sekä aika. Kuutiossa näkyvät arvot ovat liike- ja tuotekohtaiset myyntimäärät. Kuviosta nähdään, että tuotetta 1 on myyty ensimmäisellä vuosineljänneksellä 20 kpl, joista 5kpl myytiin liikkeessä 1 ja 15kpl myytiin liikkeessä 2. Tuotetta 1 myytiin koko vuonna kaikkiaan 60kpl. Kaikkia kolmea tuotetta myytiin koko vuonna 149 kpl.

Dimensioiden määrä tulee pitää kohtuuden rajoissa, näin säilytetään kuution helppo luettavuus. Mikäli dimensioita laitetaan 10 tai enemmän, voi kuution luettavuus muuttua hankalaksi ja vaikeaksi käsittää. Kun dataa käsitellään ja analysoidaan, se vaatii paljon laskutoimituksia kuten summauksia, mikä vie paljon aikaa sekä prosessointi tehoja. Kuutioissa on se huomattava etu, että ne ovat valmiiksi koottuja ja summattuja tietopaketteja, joten niitä on helppoa ja nopeaa käsitellä.

4 Toteutus

4.1 Nykytilanne

TeliaSonera Finlandilla (TSF) on tällä hetkellä käytössään CVD2 – EDW -tietovarasto ja sen metadataa tarkastellaan CVD Intra – wikisovelluksella. CVD2 -tietovarastoon populoidaan metadataa yrityksen omalla Fileloader ETL -työkalulla. Fileloader on C++ kielellä kirjoitettu ohjelma. Se on tarkoitettu siirtämään dataa .flat -tiedostoilla tai .XML-tiedostoilla tietovaraston tietokantoihin. Fileloader on linkitetty käyttämään Oraclen kirjastoja ja sen ohjelmistologiikka käyttää Oraclen funktioita tietokantafunktioiden käynnistämiseen. Ohjelma ei pysty tekemään minkäänlaista prosessointia ilman käyttäjien määrittelemiä parametreja. Ohjelman ei ole tarkoitus korvata Oraclen tarjoamia työkaluja, vaan ohjelmaa käytetään kulloisenkin tilanteen vaatiman tarpeen mukaan. Joissain tapauksissa on tehokkaampaa käyttää Oraclen työkaluja. Fileloader ei sisällä graafista käyttöliittymää ja se on suunniteltu ajamaan latauksia ajastetusti käyttäjien määrittelemien parametrien mukaan. (Fileldr User's guide, TeliaSonera.)

Yritys on tulevaisuudessa siirtymässä käyttämään yhä kokonaisvaltaisemmin ETL -työkalua nimeltä Informatica PowerCenter. Informatica PowerCenter osaaminen on yleisempää ja siihen on saatavilla tukea. Järjestelmä ei siis ole niin henkilöriippuvainen kuin Fileloader, joten sen käyttöönotto ja uusien henkilöiden koulutus on huomattavasti helpompaa. Yrityksen tietovarastoinnin kannalta on tämä pitkällä tähtäimellä riskittömämpi vaihtoehto.

Fileloaderin populoima tekninen ja operatiivinen metadata on kuvattuna yrityksen CVD Intrassa metadatatauluissa. Intrassa näkyvä operatiivinen ja tekninen metadata tullaan tulevaisuudessa tuottamaan Informatica: PowerCenter ETL -välineellä. Tästä syystä yrityksellä on tarve tutkia PowerCenter ETL -välineen tuottamaa metadataa, mitä on talletettuna PWCIMATICA -repositorioon. Repositorion metadataa voidaan tarkastella Metadata Exchange (MX) Viewin tarjoamista näkymistä.

4.2 Tavoite

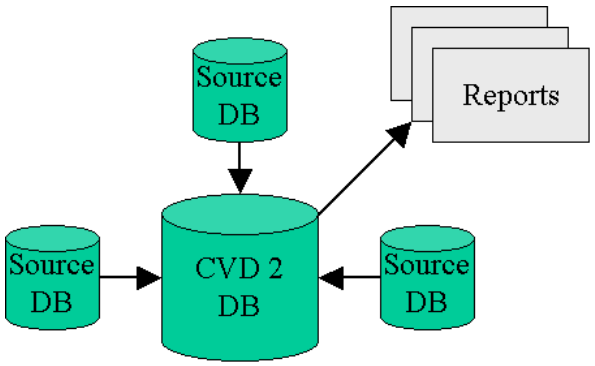
Opinnäytetyössä on tarkoitus tutustua tietovarastointi ja liiketoimintatiedon hallinta – ympäristöön ja tutkia yrityksen PWCIMATICA - repositorion metadataa. Tarkastellaan minkälaista metadataa PowerCenter ETL -väline tallentaa PWCIMATICA -repositorioon ja löytyykö sieltä metadataa, mikä vastaa yrityksen nykyisen ETL -välineen Fileloaderin tuottamaa metadataa. Tarkoitus on myös tutkia repositorion näkymiä yleisesti ja katsoa löytyykö sieltä vastaavuuksien lisäksi muuta hyödyllistä ja informatiivista metadataa, mitä kannattaisi alkaa

hyödyntämään esimerkiksi Intrassa näkyvissä lataustiedoissa. Lataustiedoista ilmenee mm. lähdetiedostojen latausprosessien tilanne. Selvityksen avulla yritys pystyy saamaan nopeammin käsityksen Informatican metadatasta ja löytää yrityksen nykyistä metadataa vastaavat metadatat. Vastaavuuksien löytäminen auttaa siirryttäessä käyttämään uutta ETL -työkalua ja tuotettaessa CVD Intran metadatataulut.

4.3 CVD2

CVD2 on TSF:n Enterprise Data Warehouse (EDW) joka sisältää suuria määriä (n. 30 teratavua) dataa noin 80 :sta lähdejärjestelmästä (muut data warehouset/datamartit). Tietovaraston tekninen metadata ohjaa muunnos- ja latausprosessia (ETL), millä haetaan lähdejärjestelmistä dataa. Talletettavaa dataa on kaikki yrityksen toimintaan liittyvä tärkeä tieto mm. laskutus-, asiakas-, myynti-, televerkon konfiguraatiometadata. Dataa talletetaan, päivitetään ja haetaan päivittäin. CVD2:n tietokanta on relaatiotietokanta, joka pitää sisällään suuria määriä skeemoja. Skeema on taulujen joukko, mihin on koottu tiettyyn sovellukseen kuuluvat taulut. Skeeman toinen tarkoitus on taulujen tai näkymien hallinnointi. Kanta pitää myös sisällään paketteja ja proseduureja. (TeliaSonera CVD2 Application Overview Document for CVD2 System, 2009.)

Järjestelmä on tehty analysointi ja raportointi tarkoituksia varten. Suurin käyttäjäryhmä CVD2:ssa onkin Business Object (BO) käyttäjät. BO -käyttäjät yhdistävät järjestelmään (CVD2) käyttäen joko BO -sovellusta tai BO –Web Intelligence:n kautta käyttäen raportointi intranettiä. BO -käyttäjien raporttien metadata on liiketoiminta metadataa. Alla oleva kuvio 11 kuvaa data warehousen ja lähdejärjestelmien suhdetta. (TeliaSonera CVD2 Application Overview Document for CVD2 System, 2009.)



Kuvio 11: DW:n ja lähdejärjestelmien suhde (Teliasonera CVD2 Intra, 2012)

4.4 CVD Intra

CVD2 -tietovaraston sisältämä metadata julkaistaan tietovaraston käyttäjille yrityksen CVD Intrassa. Intran laajuus on noin 40000 sivua. Intraan on käyttöoikeudet raporttikehittäjillä ja tietovaraston kehitys- ja ylläpitohenkilöillä. Intrasta löytyvät muun muassa tietovaraston taulurakenne, taulujen kuvaukset, rajapintakuvaukset, tiedon lähdejärjestelmät ja latausajojen käyttämät parametrit.

Intran metadata päivittyy automaattisesti aina kun CVD2 -tietovaraston metadataa muutetaan, koska intraan luetaan metadata suoraan CVD2 -tietovaraston metadatatauluista, jotka päivitetään päivittäin. Intran avulla pyritään kuvaamaan tietovaraston tietoja ja tietovirtoja eri järjestelmien välillä. Osa rajapinnan kuvauksista on määritelty siten, että ne kertovat mihin tietovararaston taulun kenttään lähdejärjestelmän tieto tallennetaan. Tällä tavalla hallintajärjestelmällä pystytään seuraamaan tiettyjen rajapintojen tietovirtoja lähdejärjestelmästä tietovarastoon. (Arkko. 2007, 24.)

4.5 CVD2 ETL -prosessi

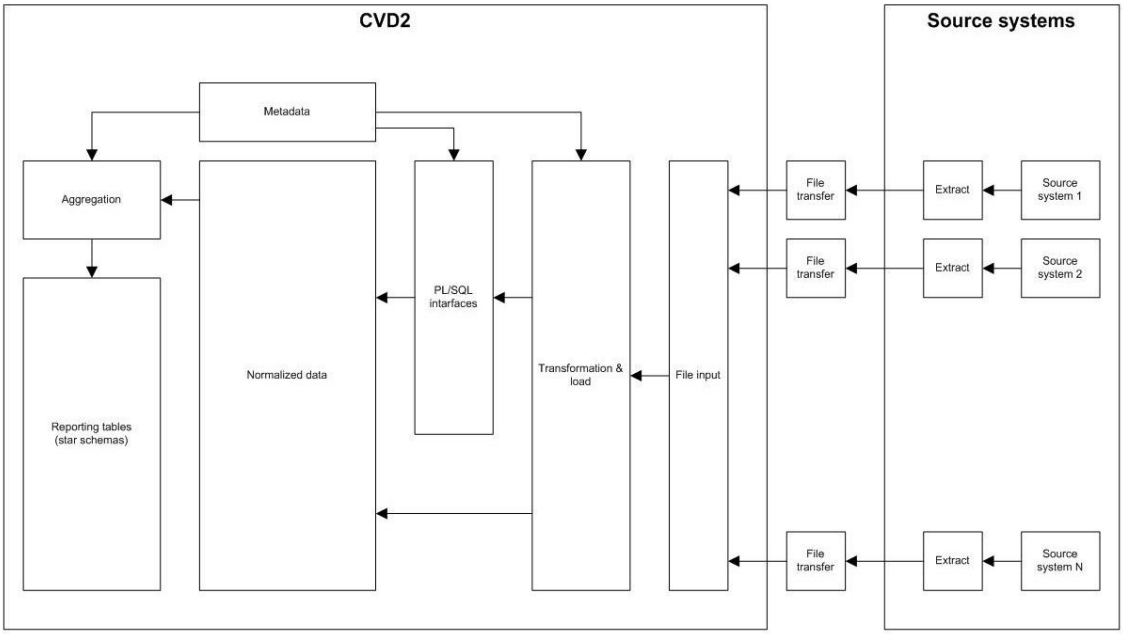
CVD2 -tietovaraston tietotaulujen tuottaminen tapahtuu Fileloader ETL -työkalulla, mikä on metadata ohjattu. Metadata ohjattu tarkoittaa sitä, että prosessin eri tehtävät suoritetaan metadatan määrityksien mukaan. Metadata ohjaa Fileloaderia ja varmistaa oikeat riippuvuus järjestykset eli toisinsanoin varmistaa, että latauksen yksittäiset proseduurit suoritetaan oikeassa järjestyksessä. Proseduurit on tehtävä tietyssä järjestyksessä, sillä ne voivat tarvita tietoja edellisistä vaiheista. Fileloader tuottaa metadataa prosessien aikana ja myös hyödyntää omaa metadataansa prosesseissa. Yksi metadataohjauksen tarkoituksista on varmistaa datan eheys ja oikeellisuus.

Yrityksellä on käytössä ELT -malli. ELT -malli poikkeaa ETL -mallista siinä, että lähdejärjestelmien tiedot siirretään vain kerran – lähdejärjestelmistä suoraan kantaan missä toteutetaan tietojen muokkaus. Tämä on mahdollista sillä Fileloader sijaitsee fyysisesti samalla palvelimella kuin CVD2-tietovarasto.

ETL -mallissa tiedot siirretään aluksi lähdejärjestelmistä ETL -palvelimelle muokkausta varten ja lopuksi muokatut tiedot siirretään tietovaraston kantaan – tapahtuu kaksi tiedonsiirtoa. ELT -mallissa on se etu, että verkkoa ei kuormiteta turhilla siirroilla ja tiedon muokkausvaiheessa päästään hyödyntämään tietovaraston palvelimen prosessoritehoja ja valmista logiikkaa.

Prosessia ohjaa ns. kehikko. Kehikossa on ohjauslogiikkaa, joka määrittelee mitä ETL-prosessissa tapahtuu. Määriteltäviä asioita on mm. missä järjestyksessä ja mihin aikaan pro-

sessin toiminnot tehdään, eli ajastukset. Se myös määrittelee mitkä lähdetiedostot poimitaan, mihin muotoon ne muokataan ja mihin ne talletetaan.



Kuvio 12: CVD2 prosessikuvaus (CVD2 ETL&Maintenance Technical Metadata, 2009. 5)

Yllä olevassa kuviossa 12 on CVD2 -prosessikaavio. Kuviossa näkyy oikealla lähdejärjestelmät (Source systems), missä tapahtuu tiedon poiminta (Extract) lähdejärjestelmistä. Lähdejärjestelmistä poimitaan päivittyneet ennalta määritellyt tietokentät. Kun tiedot on poimittu, siirretään ne CVD2 palvelimen staging area:lle odottamaan jatkokäsittelyä. Tiedostot siirretään käyttäen FTP:tä (File Transfer Protocol). Seuraavaksi Fileloader käy tiedostot läpi ja tekee tarvittavat muunnokset tiedostoille. Tehtäviä muunnoksia on esimerkiksi: liiketoiminta sääntöjen lisäys, tarkistetaan datan eheys ja suoritetaan mapping eli määritellään lähdetiedostoille oikeat sijainnit kohdetauluihin. Tiedot myös normalisoidaan mikä tarkoittaa tietojen eheyttämistä, toistuvien tietokenttien poistamista. Normalisointi nopeuttaa tietotaulujen käsittelyä ja lukuaikoja, sillä samaa tietoa ei käsitellä moneen kertaan. Viimeisessä vaiheessa ladataan tiedot CVD2:n kohdetauluihin, mistä niitä voidaan käyttää yrityksen raportointiin. Viimeisessä vaiheessa suoritetaan myös ylläpitotehtävät, mitä on statistiikoiden päivitys ym.

4.5.1 Informatica

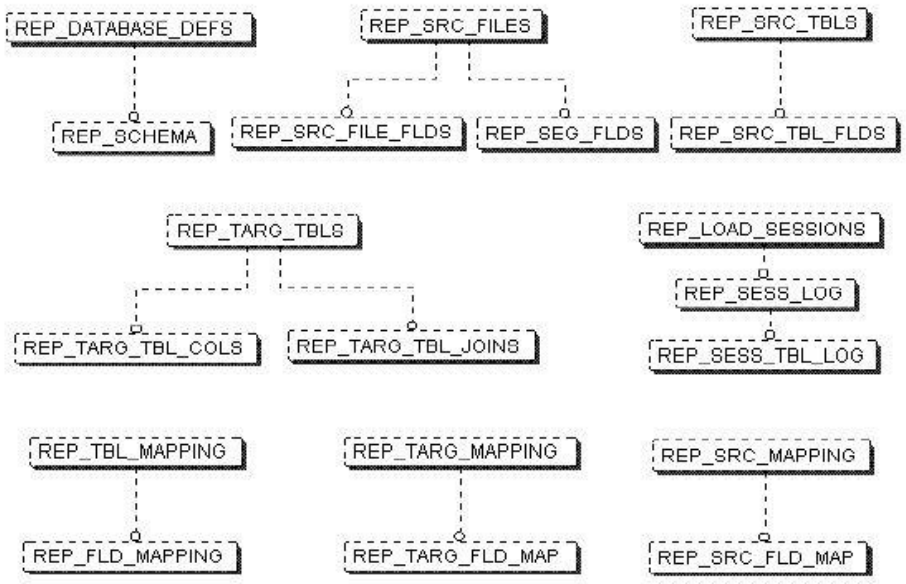
Informatica on tiedon integraatioon suuntautunut yritys, jonka mainittavin tuote on ETL-työkalu PowerCenter. PowerCenter:ssa on useampi ohjelma, mistä tärkeimpiä ovat suunnitteluun tarkoitettu – Designer sekä latauksiin ja hallintaan tarkoitetut: Workflow Manager, Workflow Monitor ja Repository Manager. Ohjelman ympäristössä voidaan poimia tietoa useista eri lähteistä, muuntaa ne haluttuun formaattiin ja viimeiseksi tallettaa haluttuun kohdesijaintiin.

Välineellä voidaan kuvata koko yrityksen metadata mikä on talletettuna yhteiseen repositorioon. Ohjelmille tuodaan metadataa paikallisesta tai yhteisestä repositoriosta. PowerCenter:lla voidaan myös tarkastella ja analysoida yrityksen liiketoimintatietoja sekä selata ja tutkia eri repositorioiden metadataa. Ohjelma tuottaa automaattisesti latauksien tilastotietoja (operatiivista metadataa) repositorioon. PowerCenter tukee OIM- sekä CWM-standardia.

4.5.2 MX Views

Informatica Metadata Exchange (MX) tarjoaa joukon relaationäkymiä, jotka mahdollistavat helpon SQL pääsyn repositorioon. Informatican Repository Manager tuottaa nämä näkymät, kun repositoriota tehdään tai päivitetään. MX Views tarjoaa informaatiota, joka auttaa analysoimaan seuraavaa metadataa: Tietokannan määrittelyt (Database definition metadata), lähde ja kohde (source and target metadata), session metadata, transformation metadata, operational metadata, user metadata. Näiden näkymien avulla saadaan helposti koostettua informaatiota repositoriosta. Jos lähdetiedoston määritelmä pitää muuttaa, niin voidaan katsoa sen sidokset, mihin muunnos tulee vaikuttamaan, näkymästä ”REP_SOURCE_MAPPING” ja ottaa ne huomioon muunnoksia tehdessä. (PowerCenter 9.1.0 Repository Guide, Informatica, 2011.)

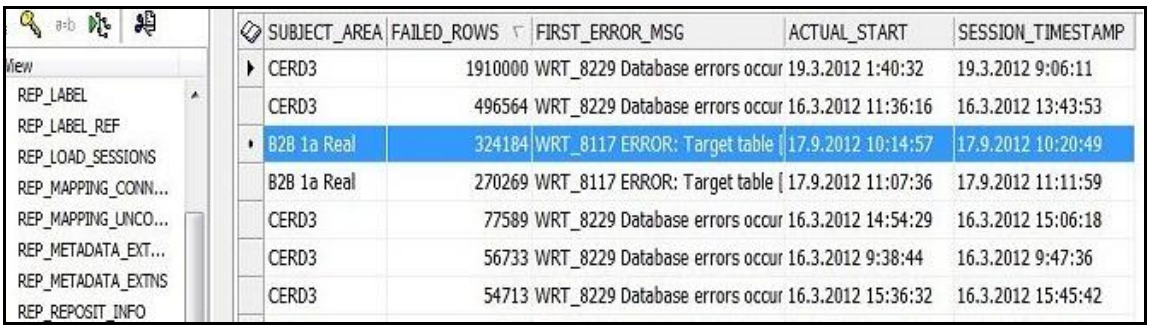
Alla oleva MX Viewin looginen malli auttaa ymmärtämään miten repositorion tauluja ja näkymiä tulisi lukea. Esimerkkinä ”REP_SESS_LOG” (sessioiden logit) taulujen kenttien kuvaukset löytyvät näkymästä ”REP_SESS_TBL_LOG”.



Kuvio 13: MX Viewin looginen malli (Repository Guide V 5.1.0, 2001, 201)

4.6 Yrityksen metadata

PWCIMATICA -repositorion metadataa tutkittiin käymällä läpi MX Viewin tauluja ja näkymiä TOAD tietokannan kehittämis- ja hallinnointiohjelmalla sekä hyödyntämällä Informatican ja yrityksen sisäisiä dokumentteja. Toad:lla voi yhdistää repositorioihin (kuva 1 alla) ja tarkastella metadataa taulutasolla (table) tai näkymätasolla (view).



Kuva 1: Toadin näkymä repositorioon. (kuva on rajattu)

Taulutasolla näkee jokaisen taulun metadatan. Näkymät ovat koosteita tauluista. Esimerkiksi Näkymä: ”REP_SOURCE_FILES”, johon on koottu kaikista tauluista lähdetiedostoihin kuuluvaa oleellista metadataa. Eli näkymien avulla voidaan tarkastella repositorion metadataa tietystä näkökulmasta. Pysyimme näkymä -tasolla ja emme lähteneet hakemaan metadatan sijainteja suoraan tauluista, sillä taulut voivat muuttua ennen kuin tämän työn tuloksia hyödynnetään. Tarkoitus oli saada yleiskuva repositorion metadatasta, selvittää minkälaista metadataa Informatica tallentaa sekä verrata sitä yrityksen nykyiseen metadataan.

Fileloaderin tuottamaa operatiivista metadataa on esimerkiksi latausprosessin ajoista syntyvät statistiikat. Alla on rajattu esimerkkikuva (kuva 2) CVD Intran prosessien lataustiedoista. Latausprosessin statiikoista poimittavaa operatiivista metadataa on mm. prosessin vaihe (processing step), viimeisin onnistunut ajo (latest success), päivityksen tiheys (frequency), viimeisimmän ajon aloitusaika (latest begin), viimeisimmän ajon lopetusaika (latest end), ajon tämänhetkinen status (status), ajon kesto (dur). Lataustietoja seuraamalla saadaan hyvä yleiskuva sen hetkisistä ajoista ja mikäli ajossa tapahtuu virhe (kuvassa statuksen kenttä: ”DEP_FAIL”) voidaan sen tapahtuman logit tarkistaa. Logeista näkyy tarkka kuvaus ajosta ja siinä tapahtuneesta virheestä. Tarkkojen lokien avulla virheiden paikannus ja korjaus käyvät huomattavasti nopeammin.

Processing step	Latest success	Frequency	Latest begin	Latest end	Status	Dur
Update f_bnum analysis tickets	03.10.2012	TODAY	04.10.2012 00:01:03		LOADING	09:14:53
NAK 2g ticket charge type	01.10.2012	DAILY_2	04.10.2012 06:09:10		LOADING	03:06:46
Tomaatti order products	03.10.2012	TODAY	04.10.2012 09:10:00	04.10.2012 09:10:10	OK	:10
Tomaatti order connections	03.10.2012	TODAY	04.10.2012 09:09:16	04.10.2012 09:09:23	OK	:07
Update f_bb_agreement_eom	03.10.2012	TODAY	04.10.2012 07:49:35	04.10.2012 08:54:52	NOT_IN_SRV	01:05:17
Datainfo invoices	29.12.2011	MONTHLY	04.10.2012 08:51:12	04.10.2012 08:51:15	OK	:03
Combined agreement agr level complete	03.10.2012	TODAY	04.10.2012 03:43:08	04.10.2012 03:43:08	DEP_FAIL	:00
Agreement upload complete	03.10.2012	TODAY	04.10.2012 03:40:33	04.10.2012 03:40:33	DEP_FAIL	:00

Kuva 2: Esimerkki lataustiedoista

CVD Intrassa on taulukuvauksia, mistä nähdään lähdejärjestelmistä ladattujen taulujen rakennekuvaukset. Rakennekuvauksista näkee kaikkien lähdejärjestelmästä poimittujen lähdetiedostojen rivien metatiedot. Metatiedot kuvaavat minkälaista tietoa rivit sisältävät. Esimerkiksi rivin kuvaus, lähdejärjestelmän antama kuvaus, onko rivi ladattu. Rakennekuvauksien avulla voidaan tarkastella minkälaista dataa taulu sisältää. Alla taulukossa 1 on esimerkki CVD Intrassa julkaistun lähdetiedoston taulurakenteesta. Taulun kentät sisältävät metadataa.

File columns

Index	Description	Source description	Loaded
1	Atlas internal product ID.	hinnasto.tuoteseq	Yes
2	Atlas fee type name.	hinnasto.hintatyyppi	Yes
3	SAP account number.	hinnasto.tulotili	Yes
4	SAP product number.	intimemuutos.intime_tuoter	Yes

Taulukko 1: Esimerkki taulukuvauksesta (kuva on rajattu).

Metadata vastaavuuksien havainnollistamista varten on tehty Excel-taulukot. Taulukoista näkee CVD Intran ja Informatica repositorion (PWCIMATICA) metadatavastaavuudet. Alla olevassa taulukossa 2 on pystysarake ”INTRA METADATA- Source Files” minkä alle on koottu CVD Intran lähdetiedostojen taulurakenteen sarakkeet”. Vieressä on pystysarake ”PWCIMATICA”, mihin on koottu vastineet Intran metadatalle. Vieressä on myös tiedon data tyyppi (Data Type) ja mistä näkymästä (view) kyseinen tieto löytyy.

INTRA METADATA – Source Files	PWCIMATICA:	**Data Type**	**View**
Properties:			
Source system	PARENT_SUBJECT_AREA	VARCHAR (240 Byte)	REP_ALL_SOURCE_FLDS
Version	SOURCE_VERSION_NUMBER	NUMBER	REP_ALL_SOURCE_FLDS
File columns:			
Column name:	**Column name:**	**Data Type**	**View**
Description	SOURCE_FIELD_BUSINESS_ NAME	VARCHAR2 (240)	REP_ALL_SOURCE_FLDS
Source description	SOURCE_FIELD_DESCRIP TION	VARCHAR (2000 Byte)	REP_ALL_SOURCE_FLDS
Loaded	Control logic		
Variable name (Fldr)	Control logic		
Data type	SOURCE_FIELD_DATATYPE	VARCHAR (40Byte)	REP_ALL_SOURCE_FLDS
Null allowed	SOURCE_FIELD_NULLTYPE	NUMBER	REP_ALL_SOURCE_FLDS

Taulukko 2: Lähdetiedostojen vastaavuuksia

Intrassa näkyvässä Properties -osiossa on taulujen perustietoa. Kaikkia Intrassa näkyviä perustietoja ei ole otettu mukaan koska osa on käsin täytettävää ja siten niille ei tarvitse löytää vastaavuuksia repositoriosta. Mukaan on otettu vain lähdejärjestelmän nimi ja versionumero Alla on taulukko 2:n kentät järjestyksessä: INTRA METADATA / PWCIMATICA – Kuvaus.

Tarkasteltu näkymä: REP_ALL_SOURCE_FLDS - Tämä näkymä näyttää kaikki lähdetiedostojen kentät ja kenttien määritykset. Näkymästä löytyi kaikki vastineet.

Properties:

- Source system / PARENT_SUBJECT_AREA - lähdejärjestelmä – Ilmaisee mistä lähdejärjestelmästä tarkasteltava taulu on peräisin.

- Version / SOURCE_VERSION_NUMBER – taulun versionumero.

File Columns:

- Description / SOURCE_FIELD_BUSINESS_NAME - Tähän kenttään on annettu informatiivinen kuvaus taulun kentästä ja sen tarkoituksesta (liiketoiminta metakuvaus). Esimerkiksi: ”Date when order was inserted into system”.

- Source description / SOURCE_FIELD_DESCRIPTION - Lähdejärjestelmän käyttämä kuvaus taulun kentälle. Esimerkiksi: ”orderDate”.
  (Intran ”Source description” tarkoittaa taulun kentän kuvausta.)

- Loaded / ohjauslogiikkaan määritelty - Onko kenttä/tieto ladattu tietokantaan. Voi olla ”yes” tai ”no”. Määritellään tiedon poimintavaiheessa tai on määriteltynä ohjaavaan logiikkaan. Kaikkia lähdetiedoston taulun kenttiä ei oteta mukaan, vaan ainoastaan tarpeelliset tietueet poimitaan.

- Variable name / ohjauslogiikkaan määritelty - Variable name on Fileloaderin käyttämä variable. Variable on ohjelmoinnissa käytettävä referenssi (arvo), millä voidaan viitata sijaintiin mihin data on tallennettu. Tässä tapauksessa variable on usein nimetty kentän arvon mukaan.

- Data type / SOURCE_FIELD_DATATYPE - Datan tyyppi. Määrittelee minkälaista tietoa kenttä sisältää, kenttä voi sisältää vain yhdenlaista tietoa. Esimerkiksi: “VARCHAR(*)”, “NUMBER”, “DATE”, “INTEGER”.

- Null allowed / SOURCE_FIELD_NULLTYPE - Määrittelee onko kentässä pakko olla arvo. Esimerkiksi: Yleisesti lomakkeissa esiintyvä kohta ”siviilisääty”: voiko siihen jättää vastaamatta eli jättää kentän tyhjäksi. Intrassa: ”yes” tai ”no”.

Latausmetadata

Latausmetadata on myös koottu samanlaiseen taulukkoon. Alla olevassa taulukossa on CVD Intrassa näkyvän lataustietojen otsikkokentät koottuna yhdeksi pystysarakkeeksi, joka on nimetty ”INTRA METADATA - Loading Info”. Vieressä on sarake ”PWCIMATICA” (Informatica repositorio) mihin on kirjattu vastineet Intran metadatalle. Vieressä on myös kentän data tyyppi (Data Type) ja mistä näkymästä (view) kyseinen kenttä löytyy. Intran kentät ”Processing step” ja ”Status” toimivat Intrassa painikkeina ja pitävät sisällään yksityiskohtaisempia tietoja. Taulukko 3:n ”Processing step” ja ”Status” kentät on aukaistu omiksi osioiksi.

INTRA METADATA - Loading Info	PWCIMATICA		
Column name	**Column name**	**Data Type**	**View**
Processing step*	WORKFLOW_NAME	VARCHAR (240 Byte)	REP_WFLOW_RUN
Latest Success	END_TIME if status ok	TIMESTAMP	REP_WORKFLOWS
Frequency	RUN_OPTIONS	INTEGER	REP_WORKFLOWS
Latest begin	START_TIME	DATE	REP_WFLOW_RUN
Latest end	END_TIME	DATE	REP_WFLOW_RUN
Status**	RUN_STATUS_CODE (6=LOADING)	NUMBER	REP_WFLOW_RUN
Duration	(END_TIME, START_TIME)	DATE	REP_WFLOW_RUN
***Processing step:**			
Status	RUN_STATUS_CODE	NUMBER	REP_WFLOW_RUN
Latest begin	START_TIME	DATE	REP_WFLOW_RUN
Latest end	END_TIME	DATE	REP_WFLOW_RUN
Dur	(END_TIME, START_TIME)	DATE	REP_WFLOW_RUN
****Status:**			
Time Stamp	END_TIME	TIMESTAMP	REP_WORKFLOWS
Area	SUBJECT_AREA	VARCHAR (240 Byte)	REP_WFLOW_RUN
Category (transfer)	Control logic		
Priority	Control logic		
Command	Control logic		
Error Message	SUBJECT_AREA (+ERROR_MSG_CODE)	VARCHAR (240 Byte) NUMBER	REP_WFLOW_RUN REP_WFLOW_RUN
Error Message Output	RUN_ERR_MSG	VARCHAR (2000 Byte)	REP_WFLOW_RUN

Taulukko 3: Lataustietojen vastaavuudet

Alla on taulukko 3:n INTRAMETADATA ja PWCIMATICAN kentät järjestyksessä kuvauksineen.

Näkymät:

REP_WFLOW_RUN - tämä näkymä näyttää kaikkien latauksien statistiikat.

REP_WORKFLOWS – tämä näkymä sisältää informaatiota yksittäisistä prosessin vaiheista ja ajastuksista.

Column name:

- Processing step / WORKFLOW_NAME - prosessin vaihe. Lähdetiedostot ladataan tietyssä järjestyksessä, sillä ne voivat tarvita tietoja edellisistä vaiheista (summien laskeminen).

- Latest success / END_TIME if load status ok – viimeisin onnistunut lataus.

- Frequency / RUN_OPTIONS – määrittelee päivitystiheyden ja tavan.

PWCIMATICA RUN_OPTIONS määritykset:

1 = Run on demand.
2 = Run once.
4 = Run every DELTA_VALUE seconds.
8 = Customized repeat.
16 = Run on Integration Service initialization.
18 = Run on Integration Service initialization and run once.
20 = Run on Integration Service initialization and every DELTA_VALUE seconds.
24 = Run on Integration Service initialization and customized repeat.
32 = Run continuously.

(PowerCenter 9.10 Repository Guide, 180)

- Latest begin / START_TIME – latauksen aloitusaika.

- Latest end / END_TIME - latauksen lopetusaika.

- Status / RUN_STATUS_CODE - määrittelee latauksen tilan.

PWCIMATICA RUN_STATUS_CODE määritykset:	
1 = Succeeded	2 = Disabled
3 = Failed	4 = Stopped
5 = Aborted	6 = Running
15 = Terminated	

(PowerCenter 9.10 Repository Guide, 186)

- Duration / (END_TIME -START_TIME) – latauksen kesto.

Processing step:

- Status / RUN_STATUS_CODE - määrittelee latauksen tilan.
- Latest begin / START_TIME - latauksen aloitusaika.
- Latest end / END_TIME - latauksen lopetusaika.
- Dur / (END_TIME -START_TIME) – latauksen kesto.

Status:

- Time Stamp / END_TIME – virheen/tapahtuman ajankohta.
- Area / SUBJECT_AREA - virheviestiin liittyvä lähdejärjestelmä.
- Category / Ohjauslogiikasta - prosessin tyyppi. Esimerkiksi: Siirto “transfer”
- Priority / Ohjauslogiikasta - prosessin Tärkeysjärjestys/tärkeysaste.
- Command / Ohjauslogiikasta – prosessin käynnistänyt komentosarja ja sen tiedosto-polku.
- Error Message / SUBJECT_AREA + ERROR_MSG_CODE – aihealue mihin on lisätty virhe-viestin tunnus. Aihealue on usein lähdejärjestelmän nimi. Informatican suhteen usein

myös puhutaan kansiosta (Folder) eli mihin lähdejärjestelmän tiedostot on fyysisesti tallennettu.

- Error Message Output / RUN_ERR_MSG – yksityiskohtainen virheviesti.

Ehdotuksia CVD-Intran metadataan

PWCIMATICA:sta löytyi paljon operatiivista metadataa, mikä voisi olla hyödyllistä sekä mielenkiintoista latauksien seurannan ja virheiden selvittämisen suhteen. Sieltä on mahdollista saada paljon yksityiskohtaisempaa tietoa menevistä latauksista ja niiden statistiikasta kuin yrityksen Intrassa tällä hetkellä on esillä. Ehdotuksien käyttöön ottaminen vaatisi joidenkin Intran näkymien uudelleen suunnittelua. Sinne tulisi lukuisia uusia kenttiä, missä näkyisi tarkempaa statistiikkaa latauksista. Alla olevassa taulukossa 4 on kooste operatiivisesta metadatasta selityksineen.

Metadata	**View**	**Kuvaus**
SESSION_NAME	REP_SESS_LOG	Session nimi
SESSION_ID NUMBER	REP_SESS_LOG	Session ID
SESSION_INSTANCE_NAME	REP_SESS_LOG	Session instanssin nimi
SUCCESSFUL_ROWS	REP_SESS_LOG	Onnistuneesti kirjoitettujen Target rivien määrä
FAILED_ROWS	REP_SESS_LOG	Target virherivien määrä
SUCCESS-FUL_SOURCE_ROWS	REP_SESS_LOG	Onnistuneesti kirjoitettujen lähderivien määrä
FAILED_SOURCE_ROWS	REP_SESS_LOG	Lähteen virherivien määrä
FIRST_ERROR_CODE	REP_SESS_LOG	Ensimmäinen virhekoodi
FIRST_ERROR_MSG	REP_SESS_LOG	Ensimmäinen virheviesti
LAST_ERROR_CODE	REP_SESS_LOG	Viimeisin virhekoodi
LAST_ERROR (last error msg)	REP_SESS_LOG	Viimeisin virheviesti
SESSION_LOG_FILE	REP_SESS_LOG	Session logitiedosto
BAD_FILE_LOCATION	REP_SESS_LOG	Hylätyn tiedoston sijainti.
LOG_FILE	REP_WFLOW_RUN	Workflow:n logitiedoston polku ja nimi
RUN_ERR_CODE	REP_WFLOW_RUN	Workflow:n virheviestin tunnus

Taulukko 4: Ehdotuksia metadataan.

PWCIMATICAN Target metadata

Tutkimme myös PWCIMATICAN kohdetauluja (targets). Kun lähdetiedostoista (Source files) on poimittu ja käsitelty halutut taulut, ne talletetaan CVD2:n kohdetauluihin (Informatica käyttää kohdetauluista nimitystä: Targets). Selvitimme minkälaista metadataa Informatica tallentaa kohdetauluista ja niiden kentistä. Target-taulujen metadataa löytyi hyvin. Alapuolella on kaksi taulukkoa (taulukot 5 ja 6) mihin on koottuna kahden näkymän metadataa. Taulutason näkymän ja kenttätason näkymän:

View: REP_TARG_TBLS

Column name	Data Type	Kuvaus
SUBJECT_AREA	VARCHAR2 (240)	kansion nimi
TABLE_NAME	VARCHAR2 (240)	Taulun nimi
BUSNAME	VARCHAR2 (240)	Taulun Business nimi
VERSION_ID	INTEGER	Kansion version ID
VERSION_NAME	VARCHAR2 (240)	Kansion version nimi
DESCRIPTION	VARCHAR2 (2000)	Taulun kuvaus
FIRST_COLUMN_ID	INTEGER	Linkki taulun ensimmäiseen kenttään
TABLE_CONSTRAINT	VARCHAR2 (2000)	Taulun sidos
CREATE_OPTIONS	VARCHAR2 (2000)	Taulun luomisvaiheen määritykset
FIRST_INDEX_ID	INTEGER	Linkki enimmäiseen indeksiin
LAST_SAVED	VARCHAR2 (30)	Viimeisin tallennusajankohta
TARGET_VERSION_NUMBER	NUMBER	Kohdetaulun versionumero
SUBJECT_ID	NUMBER	Kansion ID
TABLE_ID	NUMBER	Taulun ID.

Taulukko 5: Taulutason target metadataa.

“SUBJECT_AREA” eli aihealue viittaa lähdejärjestelmän nimeen. Informaticassa puhutaan myös kansiosta (Folder), mihin kyseisen lähdejärjestelmän tiedostot fyysisesti talletettu.

View: REP_TARG_TBL_COLS

Column name	Data Type	Kuvaus
SUBJECT_AREA	VARCHAR2 (240)	Kansion nimi
TABLE_NAME	VARCHAR2 (240)	Taulu, johon tämä kenttä kuuluu
TABLE_BUSNAME	VARCHAR2 (240)	Taulun business nimi
COLUMN_NAME	VARCHAR2 (240)	Kentän nimi
COLUMN_BUSNAME	VARCHAR2 (240)	Kentän business nimi
COLUMN_NUMBER	INTEGER	Kentän järjestysnumero
COLUMN_ID	INTEGER	Kentän ID (pääavain)
VERSION_ID	INTEGER	Kansion version ID
VERSION_NAME	VARCHAR2 (240)	Kansion version nimi
DESCRIPTION	VARCHAR2 (2000)	Kentän kuvaus
COLUMN_KEYTYPE	VARCHAR2 (50)	Kentän avaintyyppi: Pääavain, ei avain, vierasavain, pääavain & vierasavain.
DATA_TYPE	VARCHAR2 (40)	Natiivi tietokannan datatyyppi
DATA_TYPE_GROUP	CHAR (1)	Datatyypin ryhmä: C = Character D = Date N = Numeric
DATA_PRECISION	INTEGER	Desimaalinen tarkkuus numeerisille kentille tai kentän pituus CHAR -kentille
DATA_SCALE	INTEGER	Desimaalinen tarkkuus numeerisille kentille
NEXT_COLUMN_ID	INTEGER	Linkki seuraavaan kenttään
IS_NULLABLE	INTEGER	Onko NULL hyväksytty
SOURCE_COLUMN_ID	INTEGER	Linkki kentän luomislähteeseen
TARGET_ VERSION_NUMBER	NUMBER	Kohteen versio numero

Taulukko 6: Kenttätason target metadataa.

5 Johtopäätökset

Työn tavoitteena oli selvittää, löytyykö PWCIMATICA -repositoriosta samanlaista metadataa kuin yrityksen ETL -väline Fileloader tuottaa. Repositoriosta löytyi hyvin Fileloaderin tuottaman metadatan kaltaista metadataa. PWCIMATICA -repositoriosta on löydettävissä ja tuotavissa kaikki vastaava metadata kuin CVD Intrassa on tällä hetkellä esillä. Täytyy muistaa, että osa Intran metadatasta on tuotettu käsin. Informatican oma dokumentaatio on nykyisin hyvin kattava repositorion suhteen, ja MX Viewin tarjoamat näkymät helpottavat huomattavasti repositorion tutkimista.

Prosessien statistiikka-metadataa löytyy hyvin repositoriosta, ja se on hyvinkin informatiivista ja siten hyödyllistä. Yritys voi halutessaan tuoda repositoriosta nykyistä enemmän statiikka-aiheista metadataa CVD Intraan. Se tarjoaisi entistä laajemmat tiedot latausprosessien seurantaan ja vikatilanteiden selvittämiseen ja korjaukseen. Lisättävää olisi muun muassa virherivien määrät ja tarkemmat virheilmoitukset sekä polut lokeihin. Repositorio tarjoaa myös kattavat tiedot lähdemetadatasta sekä kohdemetadatasta eli nykyisin CVD Intrassa näkyvät lähdetiedostojen rakennekuvaukset ja kohdetaulujen kuvaukset voidaan tehdä PowerCenterin tuottamalla metadatalla.

Kohdeyrityksessä voisi kiinnittää enemmän huomiota metadatan ja järjestelmien dokumentointiin. Se auttaisi uusia henkilöitä järjestelmiin ja toimintatapoihin perehtymisessä. Erityistä huomiota voisi kiinnittää liiketoiminta metadataan, mitä oli yrityksen Intrassa vähän.

Tietovarastointi ja liiketoimintatiedon hallinta on käsitteeltään laaja ja pitää sisällään monia tehtäviä. DW&BI -ympäristön rakentaminen ja ylläpito isoimmissa yrityksissä onkin kokonaisen osaston vastuulla ja vaatii jatkuvaa tilannetietoisuutta koko osastolta. Tärkeää on myös nimenomaan ylläpidon yhteydenpito BI-käyttäjiin jo suunnitteluvaiheessa. Näin saadaan kaikille asianomaisille tietoisuutta aiheesta. Tietovarastointi ja liiketoimintatiedon hallinta -ympäristön antamien mahdollisuuksien ollessa tiedossa, hyötyvät yritykset sen toiminnasta todella merkittävästi ja se näkyy varmasti yrityksen tuloksessa.

Kokonaisuutena opinnäytetyöprosessi onnistui hyvin. Joitain asioita rajautui työstä pois. Liiketoimintatiedon hallintaa olisi voinut käsitellä tarkemmalla tasolla lisää. Rajaus piti yrittää pitää alusta asti mielessä, sillä sitä tehdessä oli monesti vaarana aihealueen liiallinen ”leviäminen”. Toisaalta joitain alueita olisi pitänyt tarkentaa ja käsitellä kattavammin. Paljon aikaa meni aiheeseen tutustumiseen ja käsitteiden muodostaman kokonaisuuden ymmärtämiseen.

Valittu aihe oli opinnäytetyön tekijälle täysin uusi ja siten koko opinnäytetyöprosessi oli hyvin opettavainen. Opinnäytetyöstä sai hyvän kuvan yritysmaailman tietovarastointi ja liiketoimintatiedon hallinnan – toiminnasta kokonaisuutena ja se herätti kiinnostuksen työskennellä jatkossa mahdollisesti kyseisellä puolella sekä antoi hyvät puitteet opiskella aihetta tarkemmalla tasolla lisää. Aihealueen ollessa tuttu olisi opinnäytetyössä voinut pitänyt perehtyä tarkemmin pelkästään yhteen osa-alueeseen kuten ETL-prosessiin, sen sijaan ympäristö käsiteltiin yleisellä tasolla, mutta laajasti. Kaiken kaikkiaan tämä opinnäytetyö antaa hyvän yleiskuvan tietovarastointi ja liiketoimintatiedon hallinnasta ja oli huomattava etu päästä tutustumaan toimintaan oikeassa yritysympäristössä, missä asiat näkee käytännössä.

Lähteet

Hirsjärvi, S, Remes, P & Sajavaara, P. 2008. Tutki ja kirjoita. 13.–14. osin uudistettu painos. Keuruu: Otava.

Hovi, A., Hirvonen, H. & Koistinen, H. 2009. Tietovarastot ja Business Intelligence. 1. Painos. Porvoo: WSOY.

Marco, D. & Jennings, M. 2004. Universal Meta Data Models. United States of America: John Wiley & Sons.

Marco, D. 2000. Building and Managing the Meta Data Repository – A Full Lifecycle Guide. United States of America: John Wiley & Sons.

Object Management Group. 2003. Common Warehouse Metamodel, v1.1. – CWM Chapter. Viitattu 12.10.2012. http://www.omg.org/cgi-bin/apps/doclist.pl - Hakusana: CWM

Palace, B. 1996. Data Mining. Viitattu 19.10.2012. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

Vetterli, T. 1999. A Comparison of OIM with CWM. Viitattu 12.10.2012. http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/MININGMART/PDF/oim-cwm.pdf

TeliaSonera 2012. TeliaSonera tammi–joulukuu 2011. Viitattu 16.11.2012. http://www.teliasonera.com/fi/newsroom/lehdistotiedotteet/9591/9709/teliasonera-tammijoulukuu-2011/

Julkaisemattomat lähteet:

Arkko, J. 2007. Metadatan hallinta lähdejärjestelmistä raporteille tietovarastoympäristössä. Helsingin yliopisto. Tietojenkäsittelytieteen laitos. Helsinki. Pro gradu-tutkielma.

CVD2 ETL&Maintenance Technical Metadata. 2009. TeliaSonera.

CVD Intra. 2012. TeliaSonera.

PowerCenter Repository guide V 9.1.0. 2011. Informatica.

Repository Guide V 5.1.0. 2001. Informatica

TeliaSonera CVD2 Application Overview Document for CVD2 System. 2008. TeliaSonera.

TS Metadata Requirements R001. 2012. TeliaSonera.

TeliaSonera. Fileldr User's Guide.

Kuvat

Kuviot

Taulukot